Maximizing the Value of Unstructured Sub-surface Data

The Resurgence of the Librarian in Digital Form as a Knowledge and Learning Source

Presenter: Scott Tidemann, Vice President Petrosys Middle East & Asia Pacific
Email: scott@petrosys.com.au

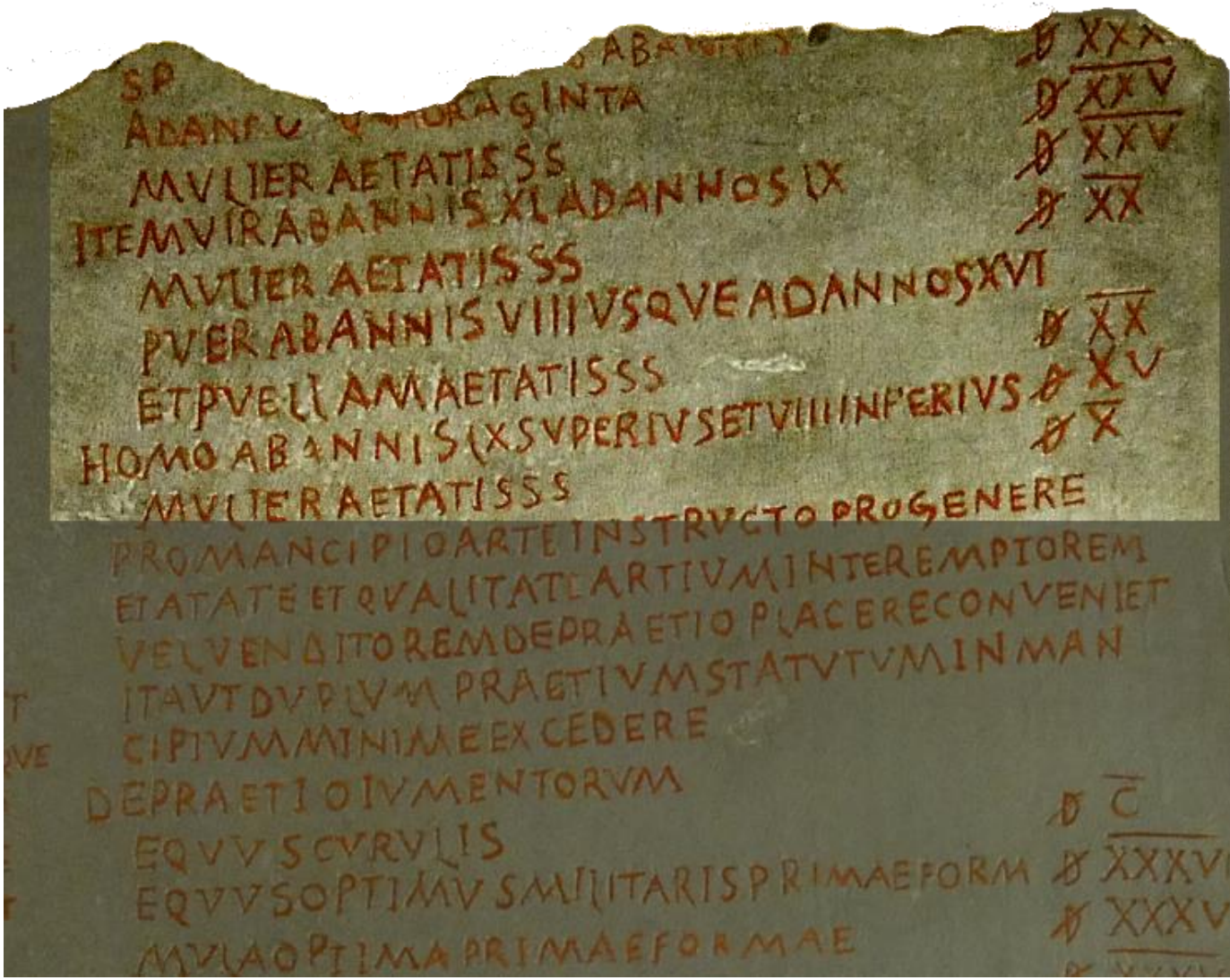
Paper authors: Dirk Adams and Volker Hirsinger, Petrosys USA, Houston, Texas.
Robert Winsloe, Digital Earth Inc., London, UK

A Brief History of the Written Word

Egyptian hieroglyphics

A Brief History of the Written Word

Roman pricelist



A Brief History of the Written Word

18th. Century
Vellum

PETROSYS

digital earth
THE NEW SEARCH FOR ENERGY

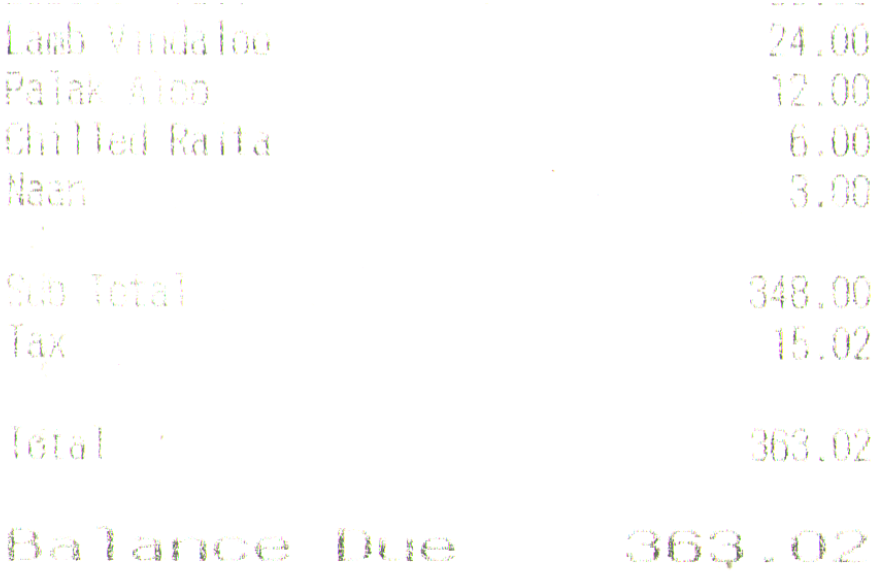
d data

A Brief History of the Written Word

20th. Century Paper

A Brief History of the Written Word

Recent High Tech Printing

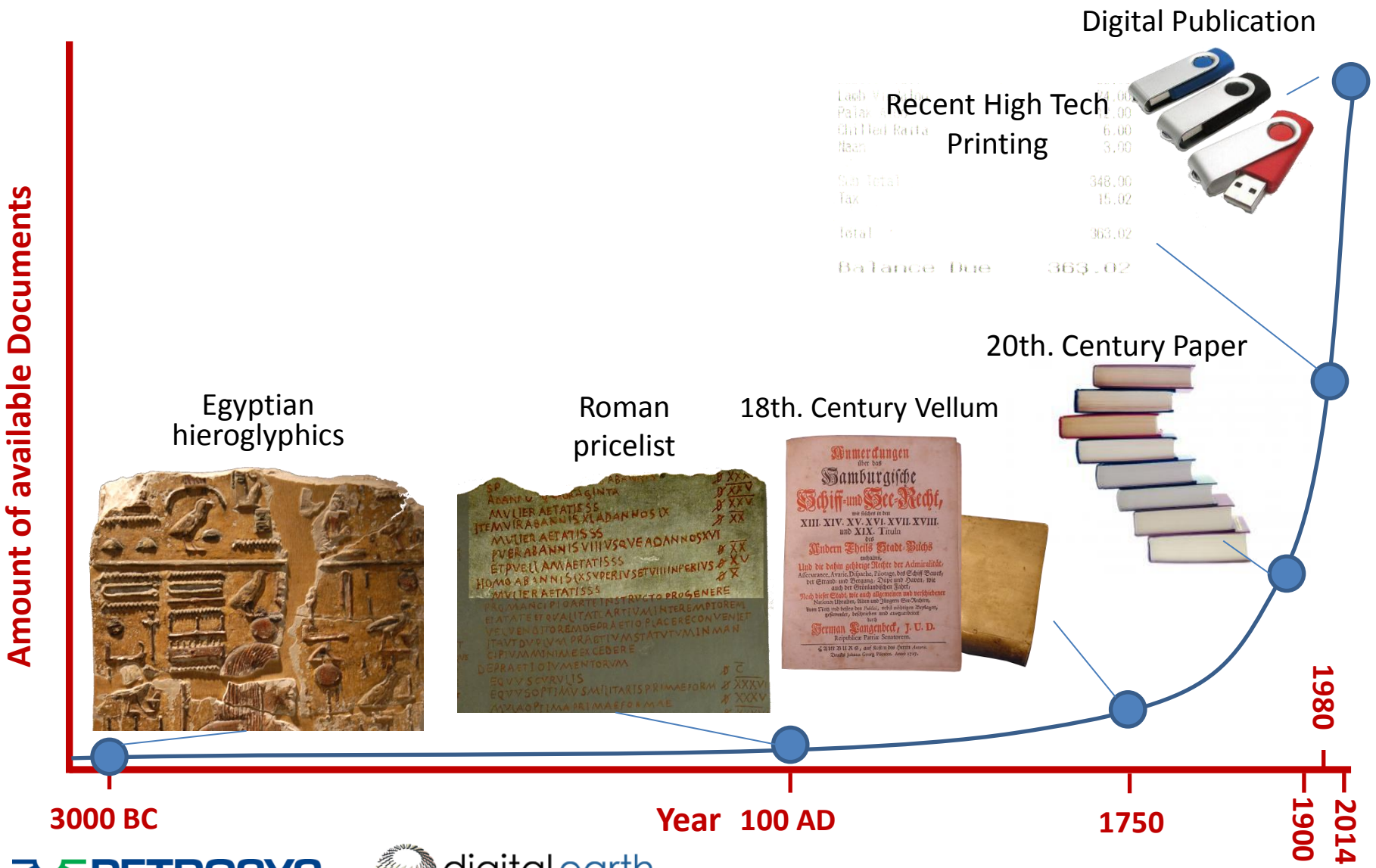


Lamb Vindaloo	24.00
Palak Aloo	12.00
Chilled Raita	6.00
Naan	3.00
Sub Total	348.00
Tax	15.02
Total	363.02
Balance Due	363.02

A Brief History of the Written Word

Digital Publication

History of the Written Word and Volume of Written Data



PETROSYS

digital earth
THE NEW SEARCH FOR ENERGY

History of the Written Word and Volume of Written Data

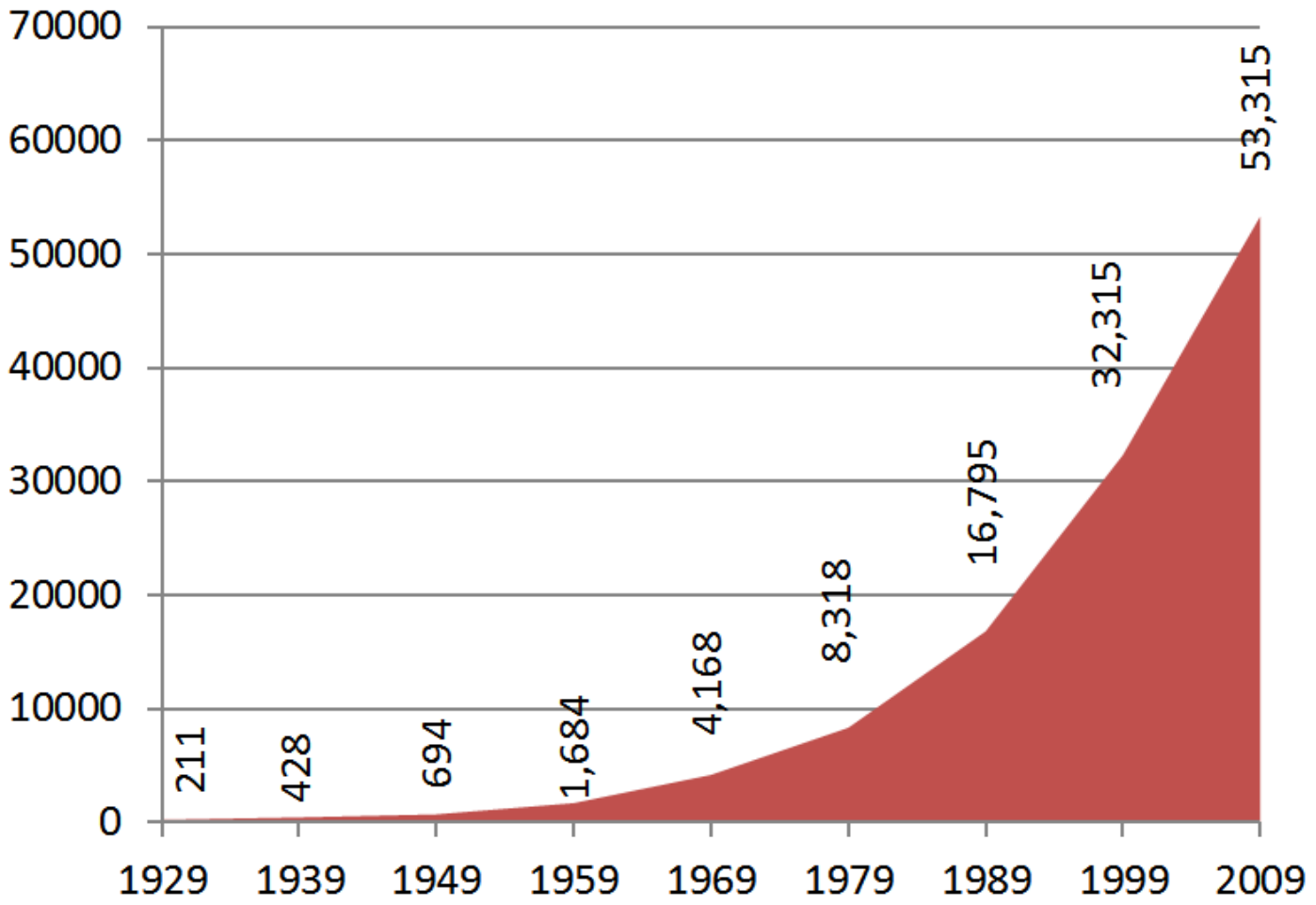
PETROSYS

digital earth
THE NEW SEARCH FOR ENERGY

The Scale of Unstructured Text in E&P

Cumulative papers published by SPE, per decade

data source: SPE Drilling & Completion



PETROSYS

digital earth
THE NEW SEARCH FOR ENERGY

The Scale of Unstructured Text in E&P

There is a huge volume of unstructured text containing E&P knowledge:

- Tabular reports – DST summaries, dumps from data stores
- Operational reports – well completions, reservoir performance, …
- Field studies
- Scientific papers
- Text content in graphic documents –
  side labels,
  well log headers, …

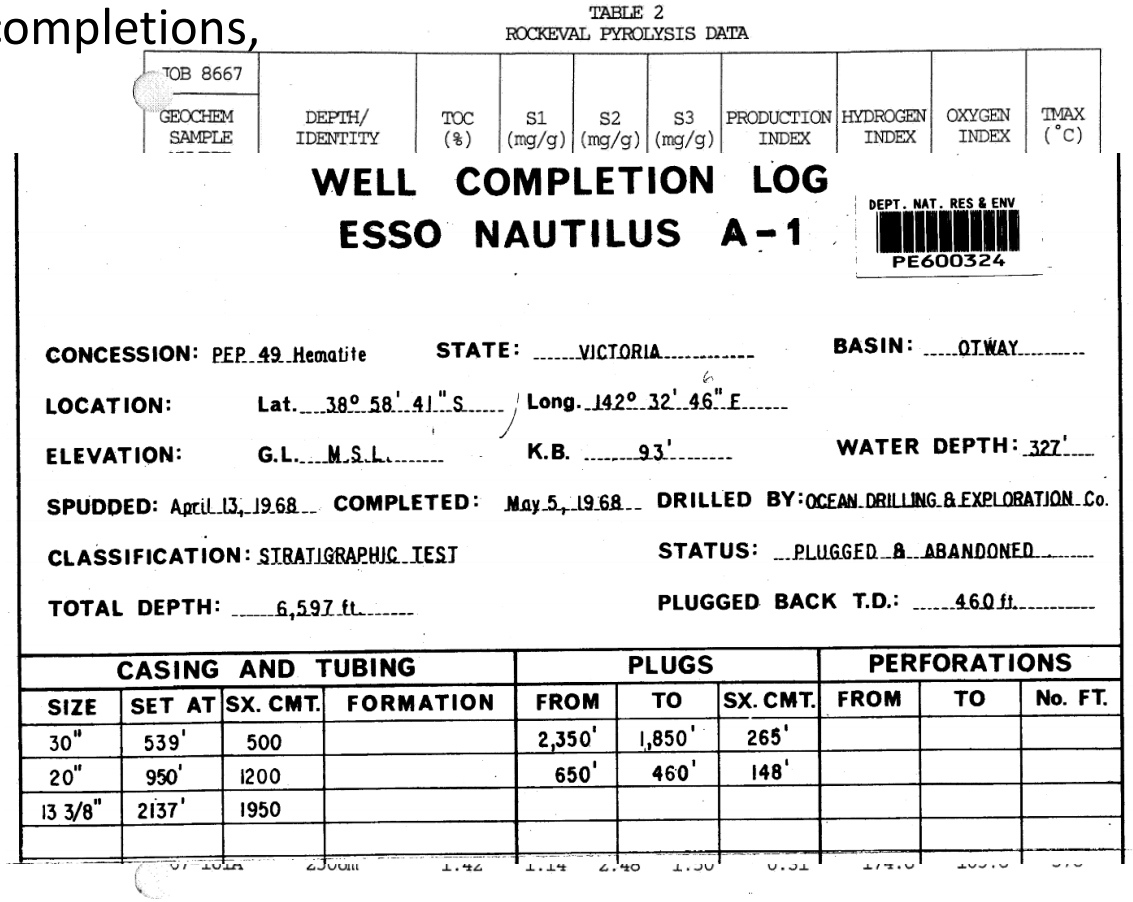


TABLE 2
ROCKEVAL PYROLYSIS DATA

TOB 8667

GEOCHEM SAMPLE	DEPTH/ IDENTITY	TOC (%)	S1 (mg/g)	S2 (mg/g)	S3 (mg/g)	PRODUCTION INDEX	HYDROGEN INDEX	OXYGEN INDEX	TMAX (°C)

WELL COMPLETION LOG
ESSO NAUTILUS A-1

DEPT. NAT. RES & ENV
PE600324

CONCESSION: PEP 49 Hematite **STATE:** VICTORIA **BASIN:** OTWAY

LOCATION: Lat. 38° 58' 41" S Long. 142° 32' 46" E

ELEVATION: G.L. M.S.L. K.B. 93' **WATER DEPTH:** 327'

SPUDDED: April 13, 1968 **COMPLETED:** May 5, 1968 **DRILLED BY:** OCEAN DRILLING & EXPLORATION Co.

CLASSIFICATION: STRATIGRAPHIC TEST **STATUS:** PLUGGED & ABANDONED

TOTAL DEPTH: 6,597 ft **PLUGGED BACK T.D.:** 460 ft.

CASING AND TUBING				PLUGS			PERFORATIONS		
SIZE	SET AT	SX. CMT.	FORMATION	FROM	TO	SX. CMT.	FROM	TO	No. FT.
30"	539'	500		2,350'	1,850'	265'			
20"	950'	1200		650'	460'	148'			
13 3/8"	2137'	1950							

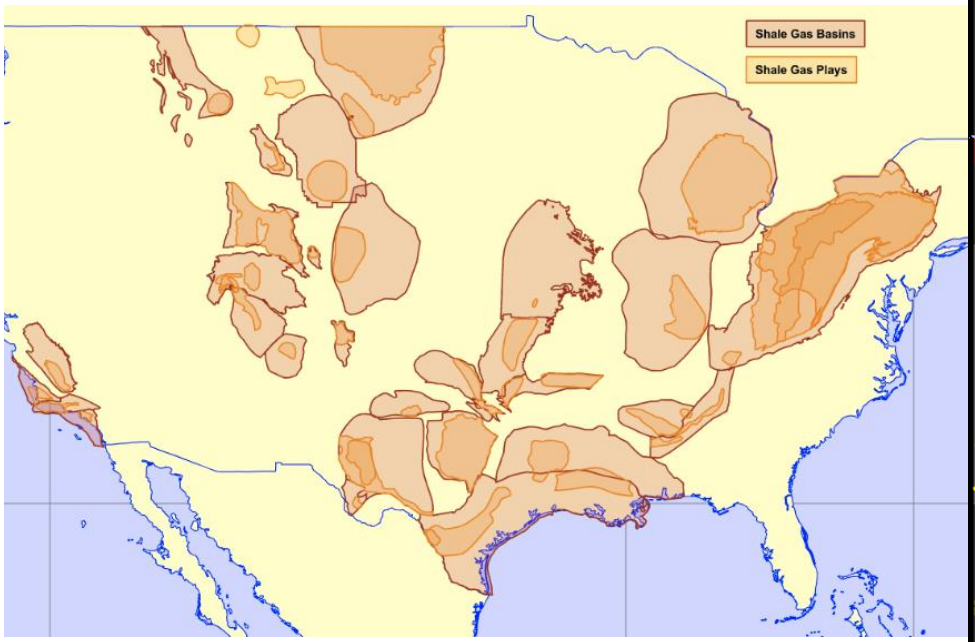
PETROSYS digital earth THE NEW SEARCH FOR ENERGY

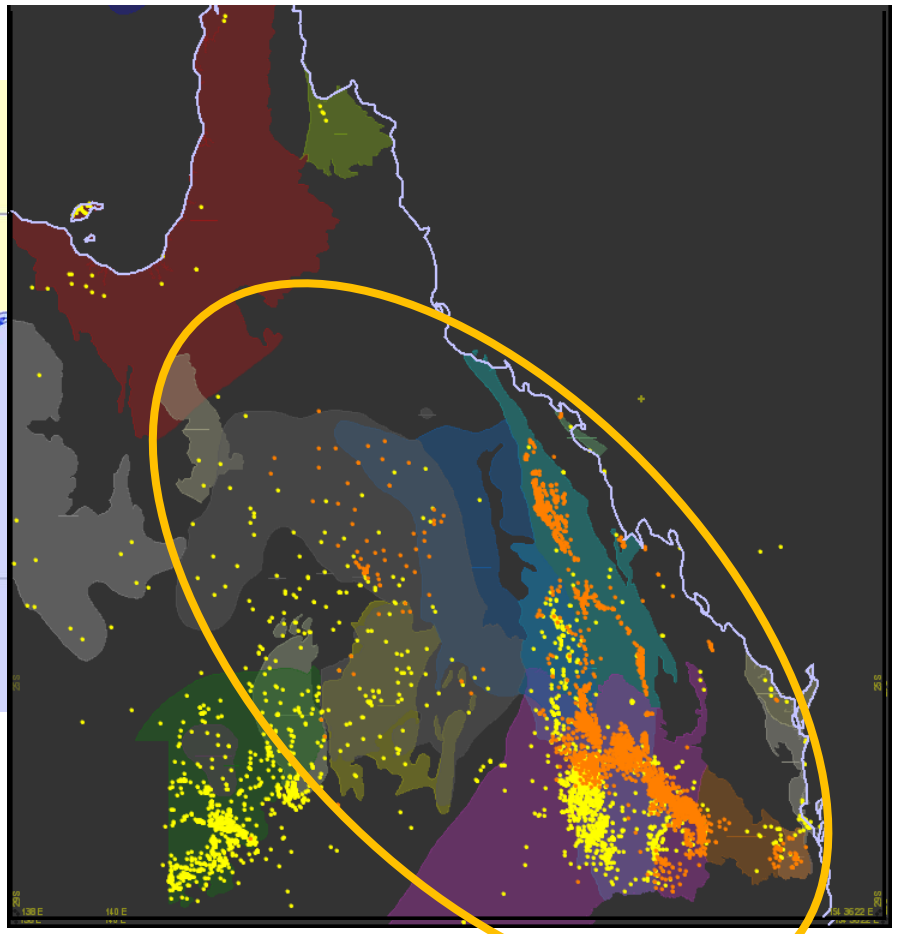
The Value of Unstructured Text to E&P

- Structured data management only addresses a tiny fraction of the wide range of knowledge acquired in petroleum E&P
- The investment in structured data management is biased towards the data that is relevant to current E&P processes
- 'Uncommercial' unstructured text has traditionally received second class treatment –
  - Geological environments that don't promise traditional plays
  - Frontier areas with high geotechnical or economic risk

The Value of Unstructured Text to E&P

- Changes in E&P technology and energy economics encourage us to take a second look at previously rejected areas
- For example, development of oil & gas from coals and shales means that we need to revisit geotechnical studies from the 50's through the 70's.





PETROSYS

digital earth
THE NEW SEARCH FOR ENERGY

The Value of Unstructured Text to E&P

- Combinations of better offshore engineering and evolving geopolitical circumstances revitalise our interest in frontier plays
- Historic reports take on new significance – discovering them becomes a challenge

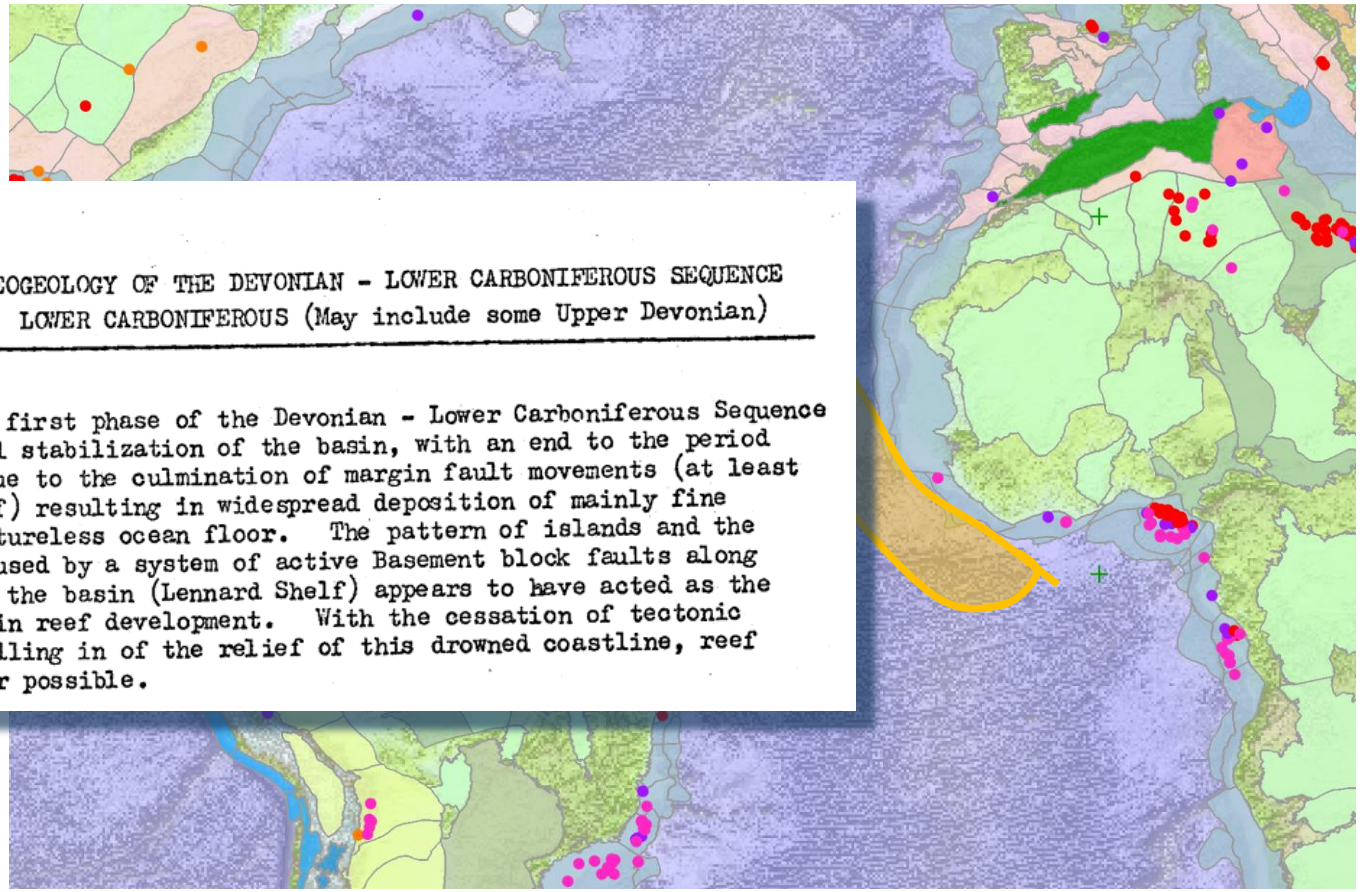


FIGURE 7. PALEOGEOLOGY OF THE DEVONIAN - LOWER CARBONIFEROUS SEQUENCE
PHASE II. LOWER CARBONIFEROUS (May include some Upper Devonian)

The end of the first phase of the Devonian - Lower Carboniferous Sequence is marked by gradual stabilization of the basin, with an end to the period of active reefing due to the culmination of margin fault movements (at least on the Lennard Shelf) resulting in widespread deposition of mainly fine clastics over a featureless ocean floor. The pattern of islands and the sea floor relief caused by a system of active Basement block faults along the north margin of the basin (Lennard Shelf) appears to have acted as the controlling factor in reef development. With the cessation of tectonic activity and the filling in of the relief of this drowned coastline, reef growth was no longer possible.

PETROSYS

digital earth
THE NEW SEARCH FOR ENERGY

Finding Knowledge In Unstructured Data

The mass of unstructured data accessible for E&P is a classical 'Big Data' issue. How do we help people find things in this?

- Library Classification Schemes
  - the Dewey Decimal (Library) System
  - More recent domain specific taxonomies
- Keyword / Subject Indexes and Thesauruses
  - Library of Congress catalogue
  - Australian AESIS Index
  - Anarchistic tags
- Natural Language Search Mechanisms
- Linking to Data Objects in a Structured Database

Classification Schemes – Library Coding Systems

Library classification systems played a key role in helping people find things in unstructured data collections, from their inception in the early 1800's until their gradual replacement with more elaborate electronic indexes at the present day -

- Dewey Dec
  - Based c
- Library of C
- ISBN numb

Holmes Principles of physical geology /

by Holmes, Arthur, 1890-1965., Holmes, Doris L., 1899-

New York : Wiley, c1978.

Subjects	● Physical geology.
Call #:	203 H73p 1978
ISBN:	0470263369
Imprint:	730 p. : ill., maps ; 24 cm.
Edition:	3d ed. / rev. by Doris L. Holmes. --
Notes:	Previous ed. published in 1965 under title: Principles of physical geology. "A Halsted Press book." Includes bibliographies and index.
OCLC No.:	ocm03705998

PETROSYS digital earth THE NEW SEARCH FOR ENERGY

Library Indexes and Taxonomies

- Most library indexing systems are based on classification schemes – taxonomies.
- The problem with defining a taxonomy is building classifications for future knowledge that has not yet been conceived.
  - For example, the Dewey Decimal Library Classification …

Dewey Decimal Classification

The Ten Main Classes

000 C
100 P
200 R
300 S
400 L
500 S
600 T
700 A
800 L
900 H

The Hundred Divisions

000 Comput
010 Bibliogr
020 Library
030 Encyclo
040 [Unassi
050 Magazi
060 Associa
070 News m
080 Quotati
090 Manusc

100 Philoso
110 Metaph
120 Epistem
130 Parapsy
140 Philoso
150 Psychol
160 Logic
170 Ethics
180 Ancient
190 Modern

200 Religior
210 Philoso

The Thousand Sections

000	Computer science, information & general works	050	General serial publications
001	Knowledge	051	Serials in American English
002	The book	052	Serials in English
003	Systems	053	Serials in other Germanic languages
004	Data processing & computer science	054	Serials in French, Occitan & Catalan
005	Computer programming, programs & data	055	In Italian, Romanian & related languages
006	Special computer methods	056	Serials in Spanish & Portuguese
007	[Unassigned]	057	Serials in Slavic languages
008	[Unassigned]	058	Serials in Scandinavian languages
009	[Unassigned]	059	Serials in other languages
010	Bibliography	060	General organizations & museum science
011	Bibliographies	061	Organizations in North America
012			
013			
014			
015			
016			
017			
018			
019			
020	Library & information sciences	070	News media, journalism & publishing
021	Library relationships	071	Newspapers in North America
022	Administration of physical plant	072	Newspapers in British Isles; in England
023	Personnel management	073	Newspapers in central Europe; in Germany

… was developed in the 19th century essentially by one man and was built on a top-down approach to classify all human knowledge, which makes it difficult to adapt to changing fields of knowledge. (Wikipedia)

PETROSYS

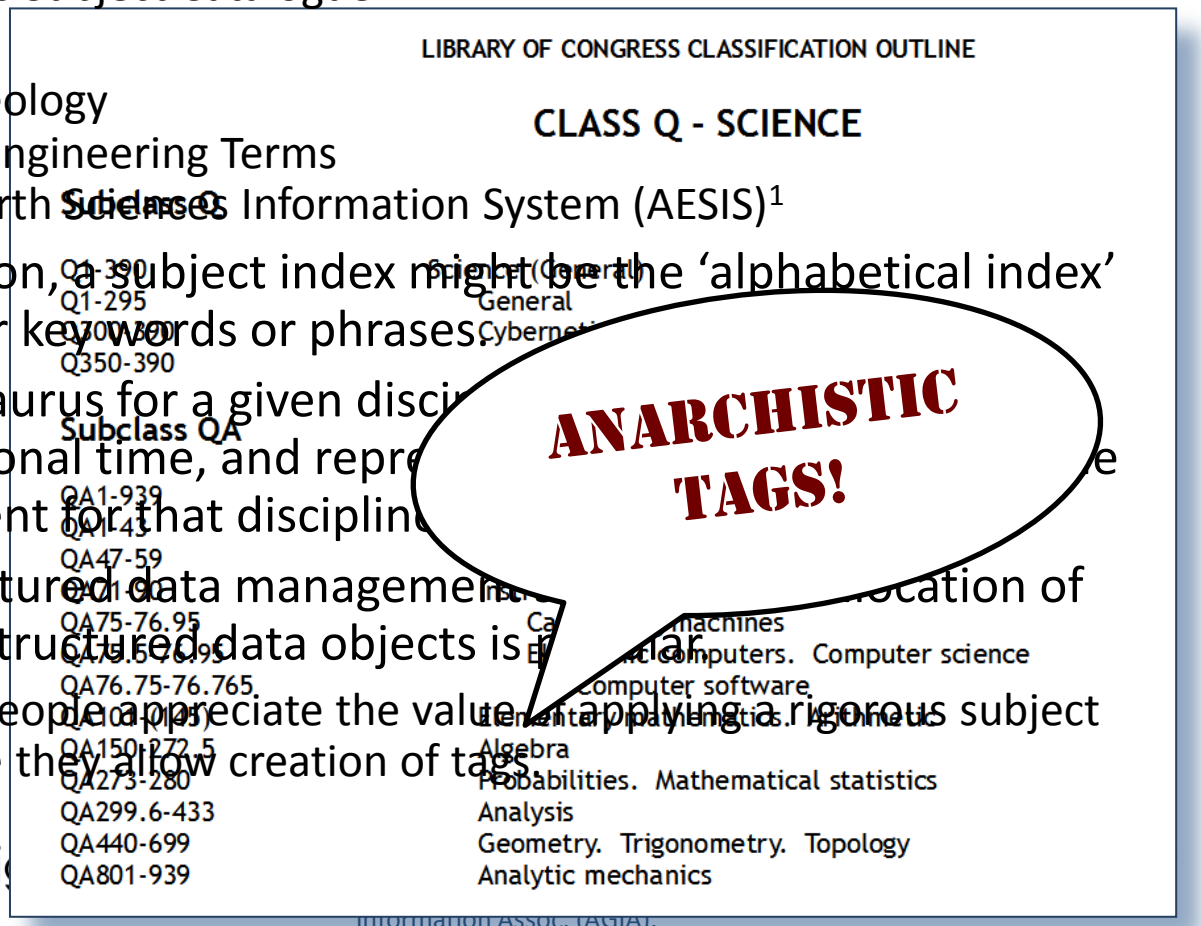
digital earth
THE NEW SEARCH FOR ENERGY

Library Indexes and Taxonomies

- Most library indexing systems are based on classification schemes – taxonomies.
- The problem with defining a taxonomy is building classifications for future knowledge that has not yet been conceived.
  - For example, the Dewey Decimal Library Classification … *was developed in the 19th century essentially by one man and was built on a top-down approach to classify all human knowledge, which makes it difficult to adapt to changing fields of knowledge*. (Wikipedia)
- Present day document indexing systems still rely largely on using taxonomies.
  - Taxonomies remain mostly time dependent. We must not lose sight of potential losses of knowledge value as we apply new taxonomies to old reports with a severe mismatch in context.

Keyword Indexes, Tags and Thesauruses

- Thesauruses and subject indexes became popular tools for librarians in the second half of the 20th century –
  - USA:
    … Petroleum Abstracts at Tulsa University
    … Library of Congress Subject Catalogue
  - Australia:
    … AGI Glossary of Geology
    … EJC Thesaurus of Engineering Terms
    … AMF Australian Earth Sciences Information System (AESIS)[1]
- In a document collection, a subject index might be the 'alphabetical index' that you can search for key words or phrases.
- The creation of a thesaurus for a given disci… investment in professional time, and repre… knowledge management for that discipline.
- In present day unstructured data management … location of 'metadata tags' to unstructured data objects is popular.
  - Unfortunately, few people appreciate the value of applying a rigorous subject index analysis before they allow creation of tags.



[1] Australian Geoscience Information Association (AGIA).

Geographic Tags and Indexes

- Most E&P data deserves to have a geographic reference.
- Geotechs and IT people like to use numeric spatial indexes. The geographic 'context' of a knowledge element is not always tied to specific coordinates –
  - Regions are often defined by context, eg. Political vs. climate vs. tectonic
  - Basin identities and outlines are interpreted
  - Field boundaries may be split by commercial and political considerations
- Our industry has invested in the creation of a range of geographical dictionaries and data sets, which are mostly proprietary, to enable consistent geographic tagging of sometimes interpretive data:
  - IHS geographic dictionary
  - Fugro
  - Government agencies, eg. MMS (USA), ANH (Colombia), GA (Australia), …
- Geographic tags and indexes remain a high value addition to the cataloguing of unstructured E&P data.

Natural Language Search Mechanisms

- Natural language search – 'like Google' – is hugely successful in helping people find things to get through their daily lives.
- Effective natural language searches are very context dependent. The success that we attribute to Google is because it is taking into account a huge range of factors that reflect a constantly changing 'knowledge' rather than the current 'truth':
  - Popular search queries and their most frequently accepted results
  - Location and other personal information about the questioner
  - Paid search preferences
  - Search criteria based on keyword and other indexes
- Natural language search may find a lot of data that is relevant, however…
  - It may also display a lot of noise
  - It will not guarantee that the searcher will see the most important data
  - Present technologies are anglo-centric, a lot of E&P literature is not in English.

Subject Indexes, Tags and Language

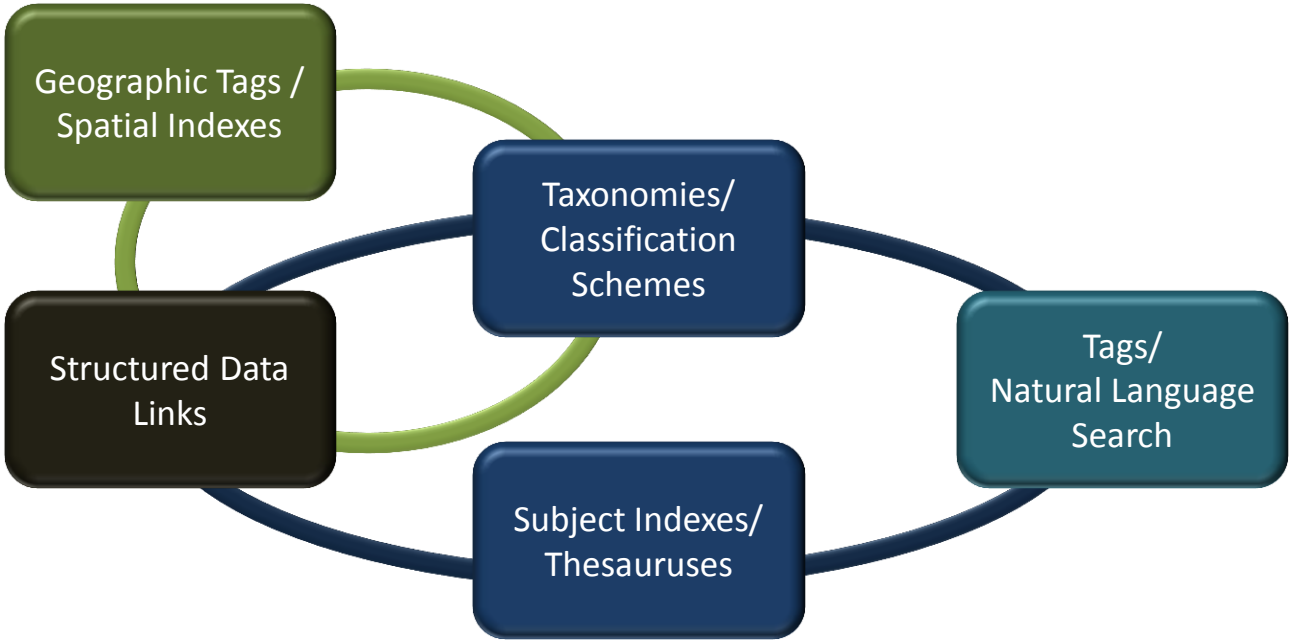
- Subject indexes, search algorithms and taxonomies have evolved with a strong bias to the English language.
- Practical application in today's industry, particularly in Latin America and Asia requires a multi-lingual approach:
  - English – TOP …. WELL COMPLETION REPORT
  - Spanish
    - Mexico – CIMA …. CARPETA DE POZOS
    - Venezuela – TOPE …. CARPETA DE POZOS
    - Argentina – TOPE …. LEGAJO
  - Chinese, Malaysian, Vietnamese, Indonesian ...

Linking Unstructured to Structured Data

- Increasingly, we have structured master data for at least the key components of our operations –
  - Well master
  - Seismic survey master
  - Production entities – fields / blocks
- The PPDM Records Management (RM) module is an effective industry standard way of storing such links in a database.
- Tools for creating the links are the software industry's challenge –
  - For new data, good data governance can promote reliable linking
  - For old data, creation of links is largely dependent on manual processing using the other tools, from natural language searches through library classifications.

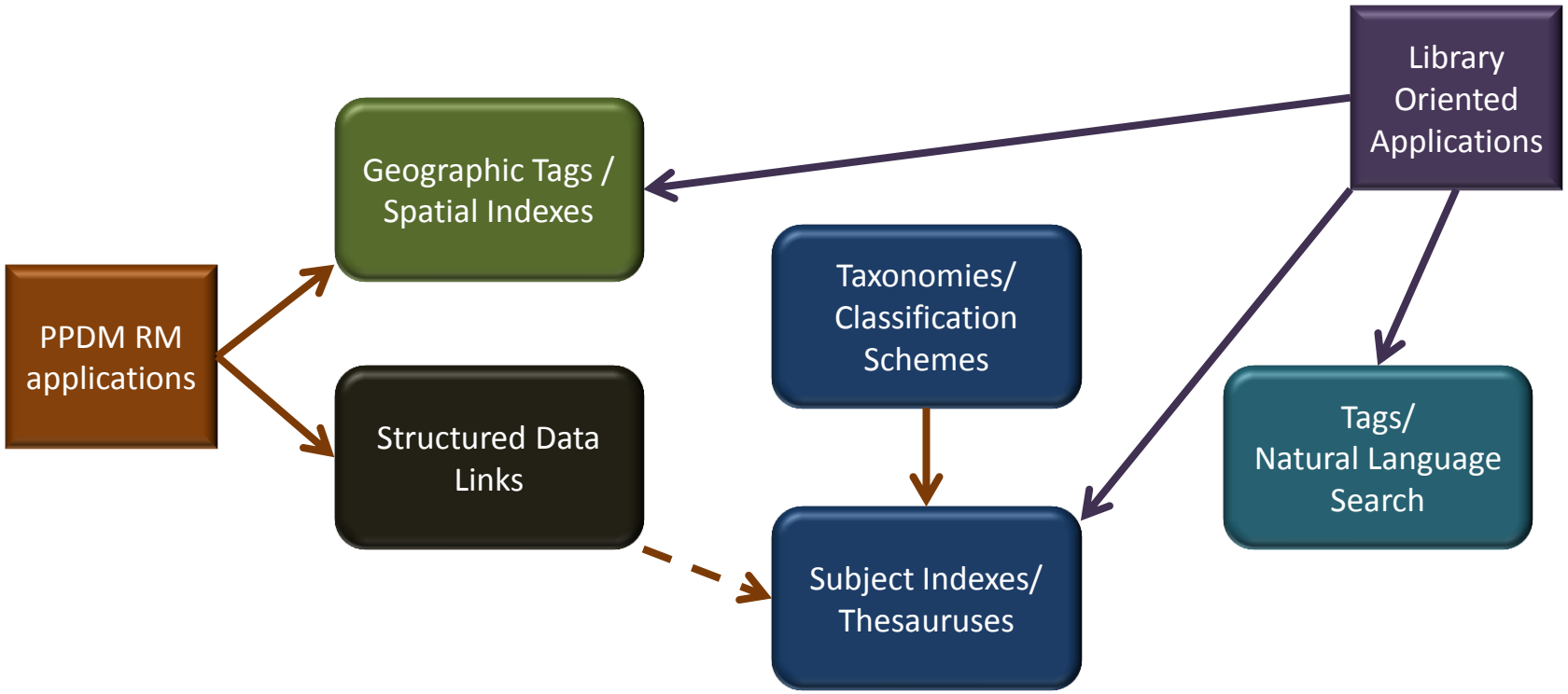
Finding Knowledge In Unstructured Data

Ideally, our unstructured data management needs to leverage all available technologies.

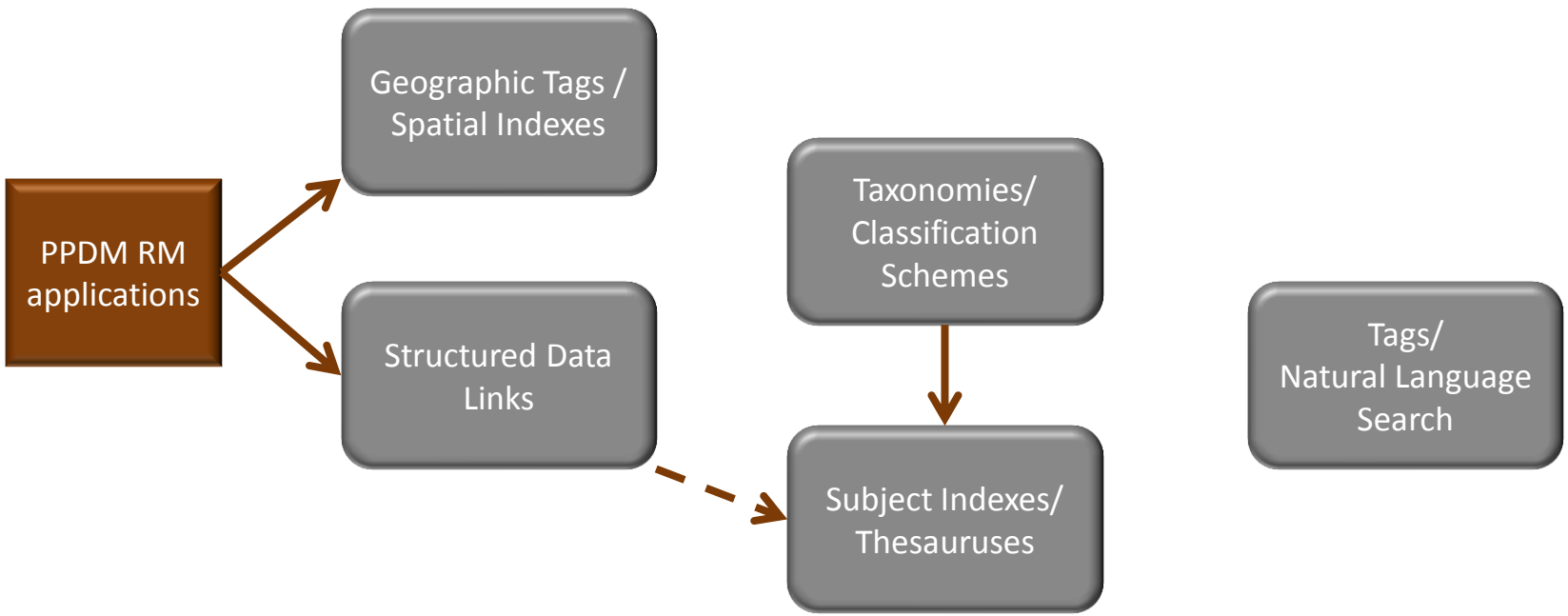


Finding Knowledge In Unstructured Data

Different applications do this in different ways, but at the end of the day they all rely on taxonomies in one way or another.

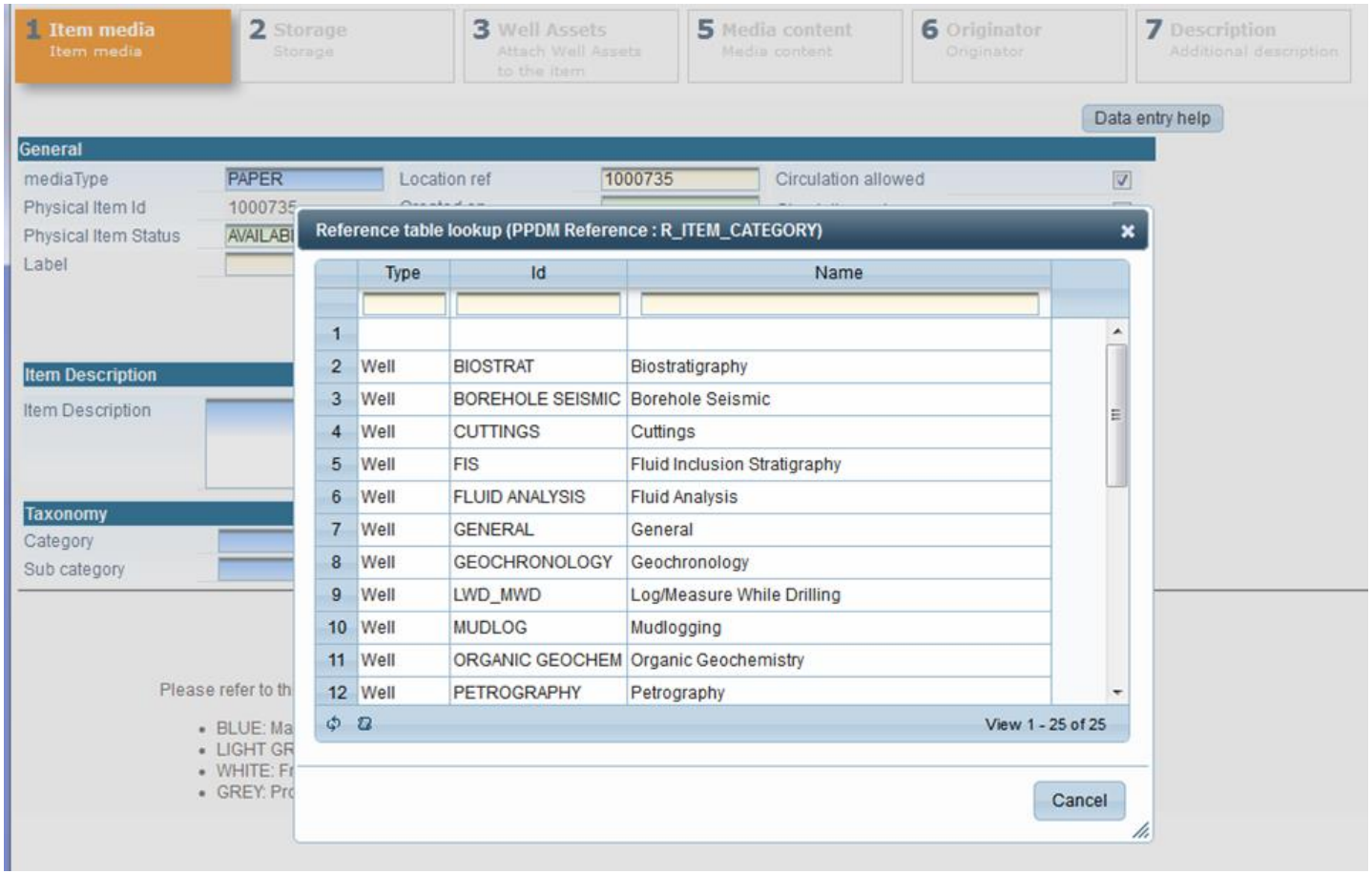


PPDM RM application example – dbMap/web



dbMap/web – A PPDM/RM approach

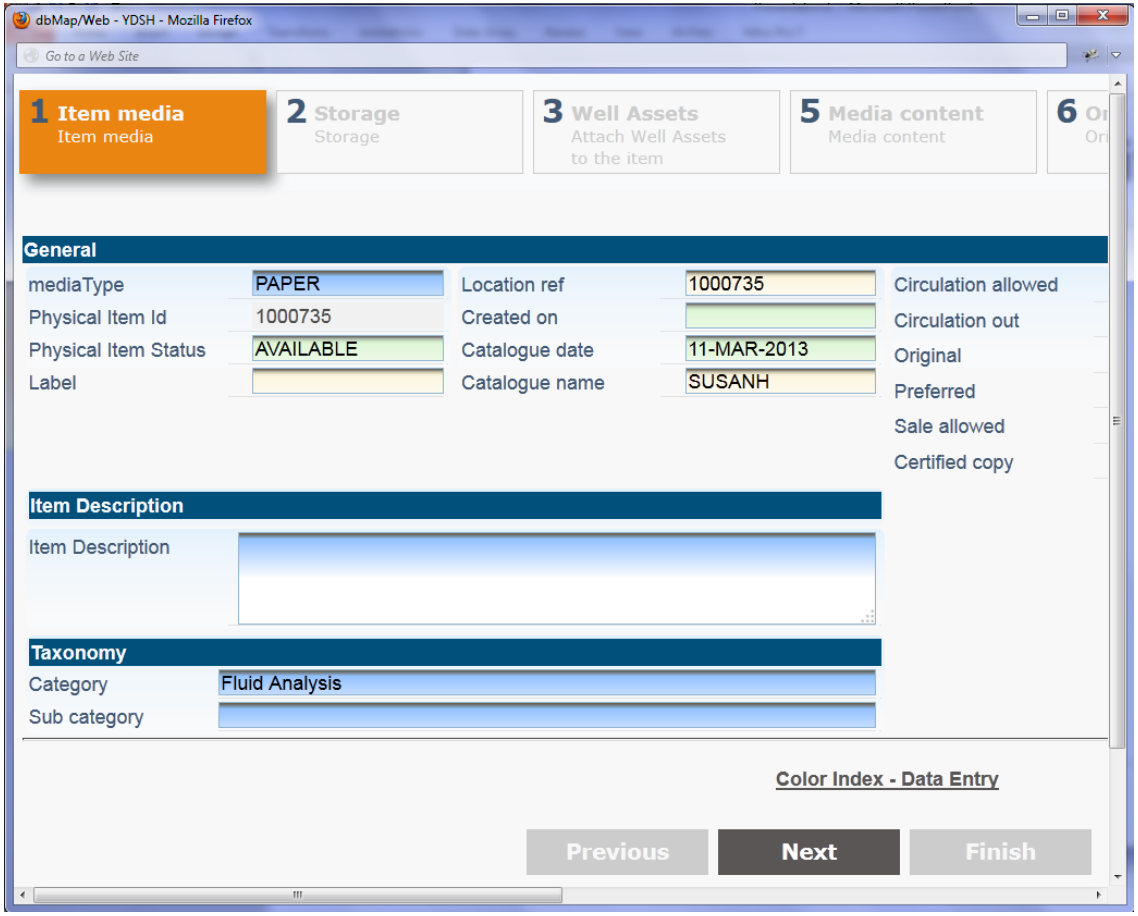
The dbMap/web unstructured data management system follows the PPDM RM model, where taxonomies are applied within the constraint of predefined tables such as R_ITEM_CATEGORY



PETROSYS digital earth THE NEW SEARCH FOR ENERGY

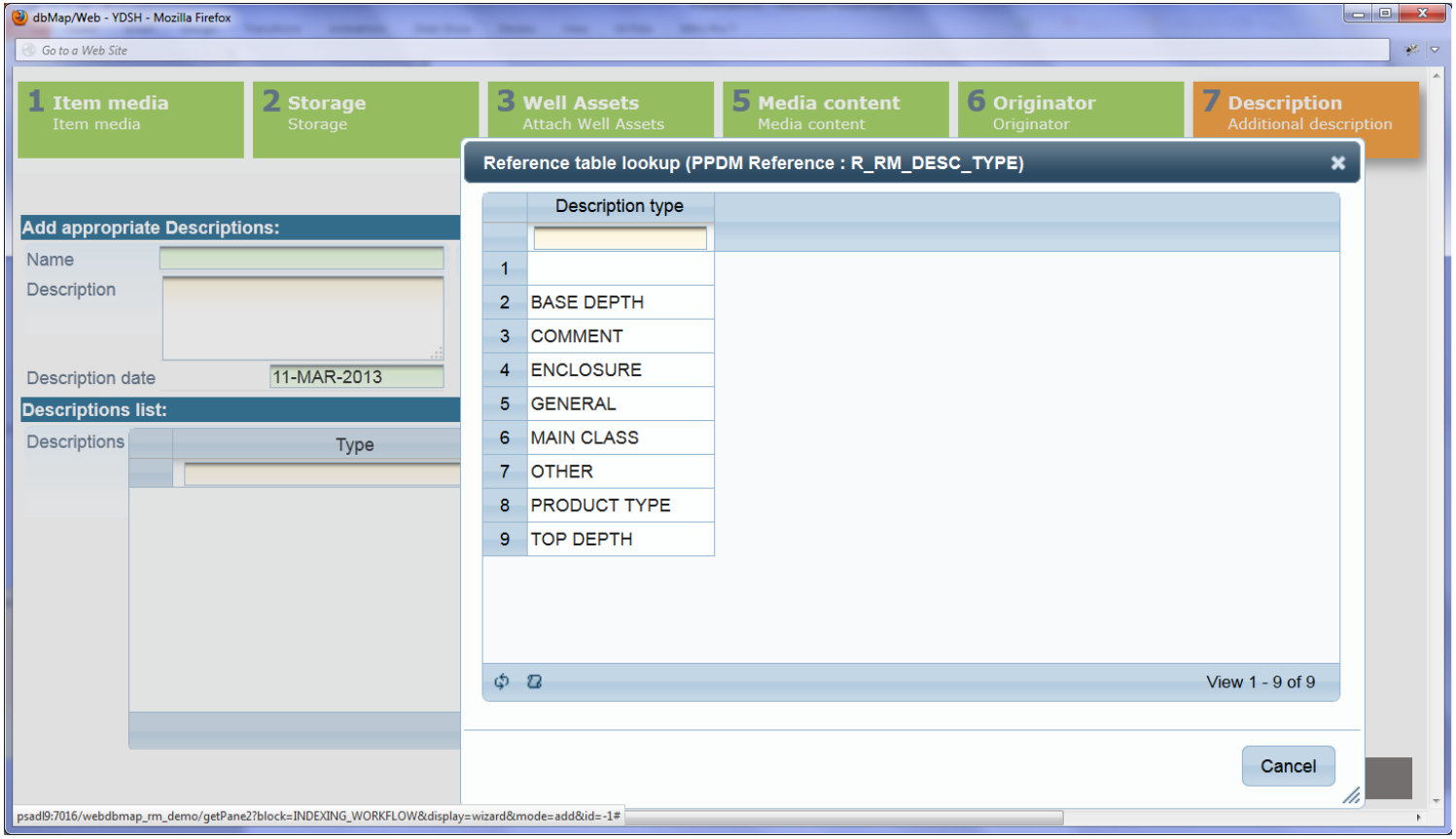
dbMap/web – A PPDM/RM approach

Each level of taxonomy in this part of the system has to be allocated it's own unique table in the PPDM schema – category vs. sub-category, for example:



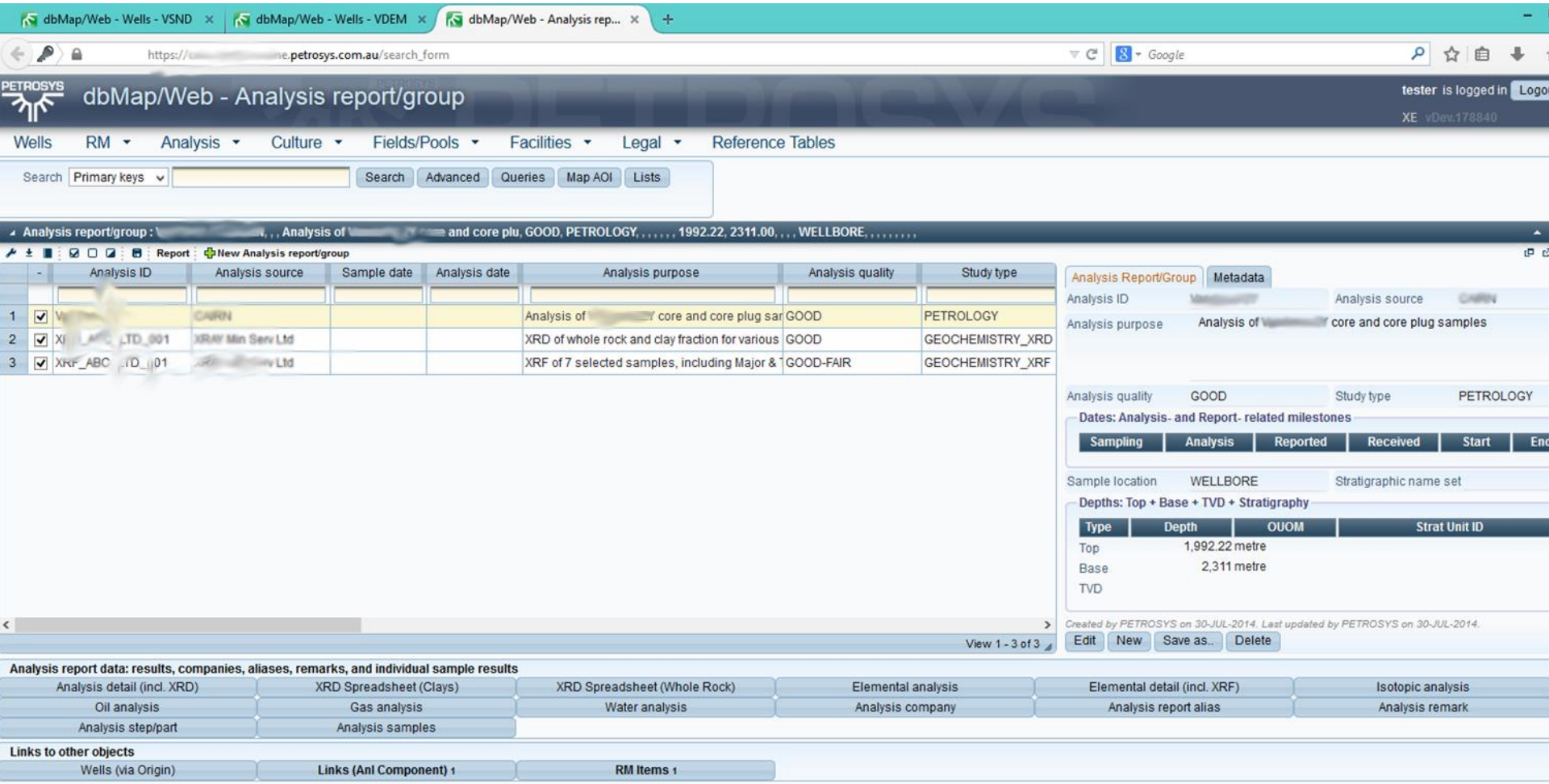
dbMap/web – A PPDM/RM approach

Addition of arbitrary additional levels of subject indices are possible using the ‘Descriptions’ on an item … what would the ideal relationship be between the intent of ‘Descriptions’ and the keyword indices of library thesauruses?

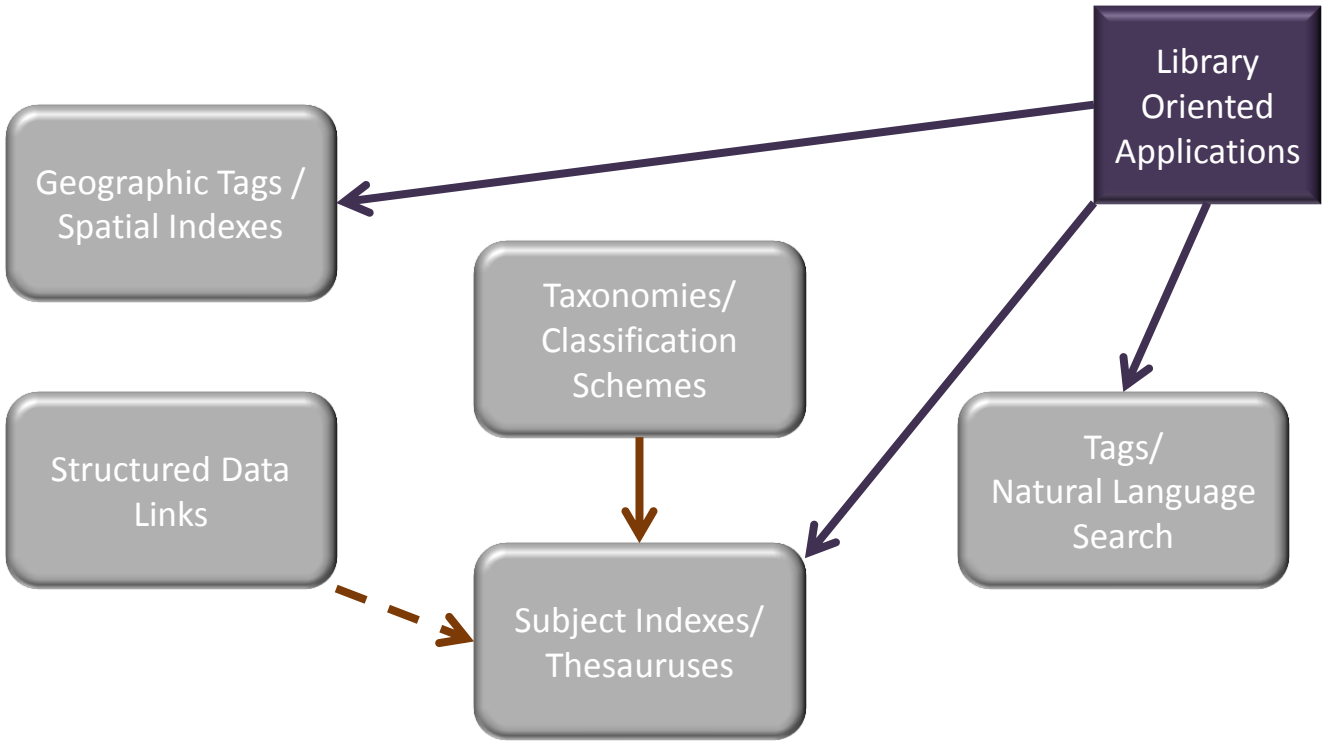


dbMap/web – A PPDM/RM approach

Below links from Core Analysis, Wells & Fields/Pools (master objects) help the user find the details for a wide range of structured & unstructured analysis.



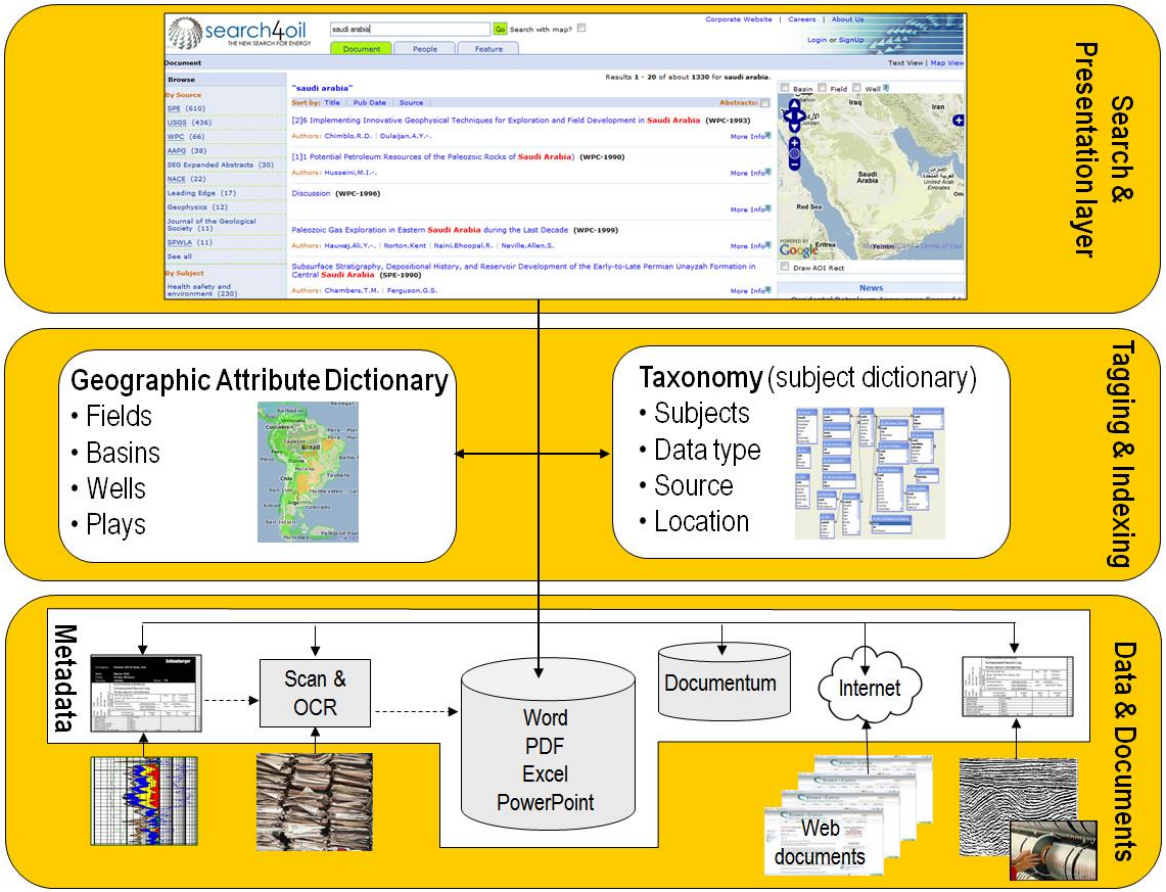
Digital Library Approach – Search4Oil



PETROSYS digital earth THE NEW SEARCH FOR ENERGY

Search4Oil – A more library centric approach

A composite approach to indexing of unstructured information is used by the Search4Oil web application -



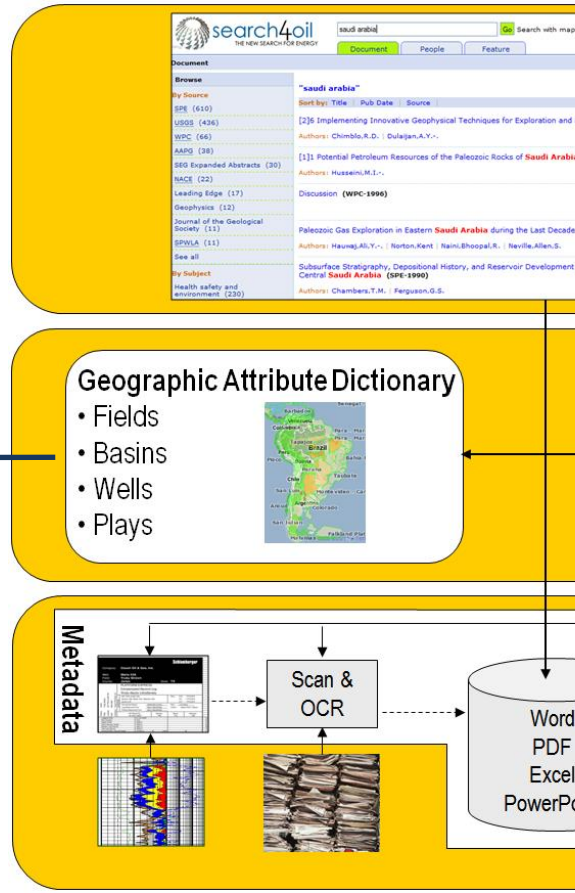
PETROSYS digital earth THE NEW SEARCH FOR ENERGY

Geographic Tags in Search4Oil

Publicly available geographic tagging has been used in Search4Oil. More sophisticated geographic tags, such as IHS, are not available to the public as they are proprietary.

Publicly available geographic tags –

Political Boundaries
… to municipality level
Basin Outlines …
Field Outlines for 18 thousand fields
Well Locations



Geographic and Free Text Searches

search4oil
THE NEW SEARCH FOR ENERGY

hydrocarbon systems Go Search with map?

Careers | About Us
Corporate Website

Document | People | Feature

Login or SignUp

Document Text View | Map View

Browse

By Source

- SPE (45)
- AAPG (10)
- Sedimentary Geology (2)
- USGS (2)
- Applied Geochemistry (1)
- Chemie der Erde - Geochemistry - Interdisciplinary Journal for Chemical Problems of the Geosciences (1)
- Journal of Geochemical Exploration (1)
- Leading Edge (1)

Results **1** - **20** of about **65** for **hydrocarbon systems** in zoom:4.

"hydrocarbon systems"

Sort by: Title | Pub Date | Source | **Abstracts:**

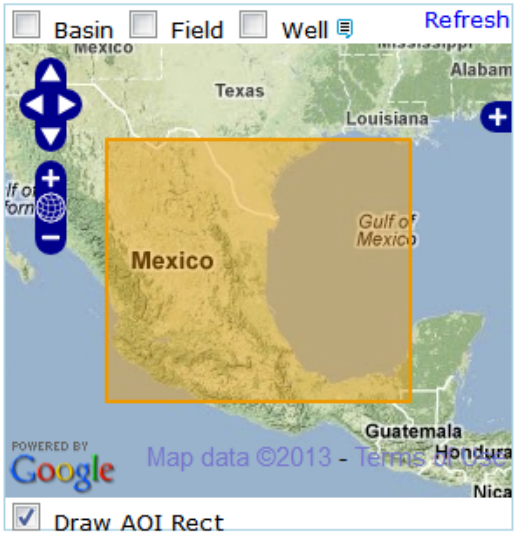
Chapter 5. Assessment of undiscovered conventional oil and gas resources-Lower Cretaceous Travis Peak and Hosston formations, Jurassic Smackover interior salt basins total petroleum **system**, in the East Texas basin and Louisiana-Mississippi salt basins **(USGS-2005)**

Authors: Dyman,T.S. | Condon,Steven More Info

Temperature Dependence of L1/L2/V Behavior in CO2/**Hydrocarbon Systems** **(SPE-1988)**

Authors: Orr,F.M. | Silva,M.K. | Larson,L.L. | Taylor,M.A. More Info

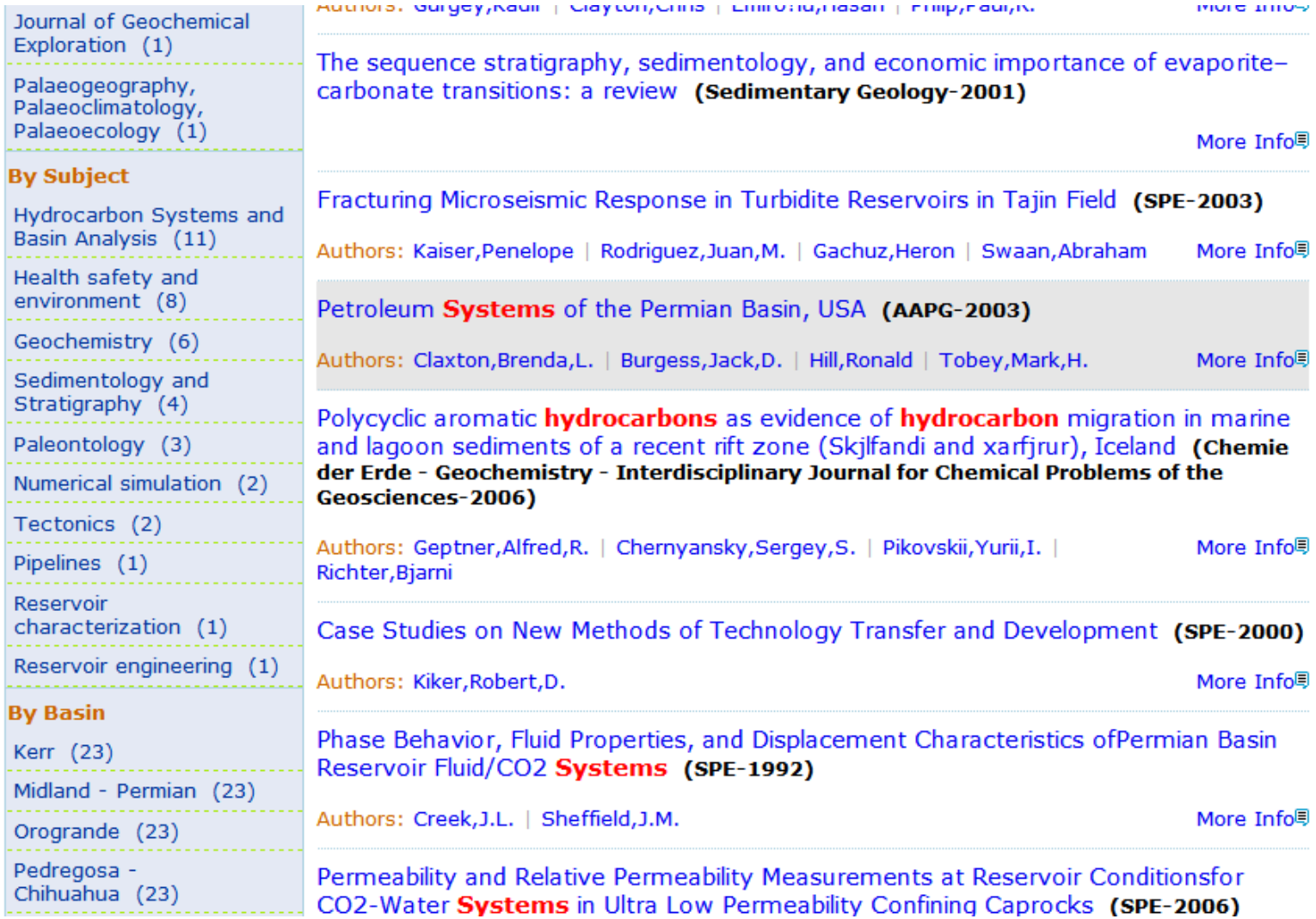
Geochemical and isotopic approach to maturity/source/mixing



PETROSYS

digital earth
THE NEW SEARCH FOR ENERGY

E&P Specific Keyword Indexes



PETROSYS digital earth THE NEW SEARCH FOR ENERGY

Native Language Search Scores

Document

External Link to the paper

Integrated Geostatistical Reservoir Characterization of Turbidite SandstoneDeposits in Chicontepec Basin, Gulf of Mexico

Related Features

Hydrocarbon Systems and Basin Analysis Petrophysics Tampico Mexico Reservoir characterization

What is this?

Authors:	Yamamto,Hiroshi, Sandria,Francisco Murguia, Abbaszadeh,Maghsood, Yazawa,Nintoku, Guerrero,Fernando Rodriguez Garza, Shimamoto,T., Takano,Osamu, Zamora,David,H.
Serial:	SPE Annual Technical Conference and Exhibition, 5-8 October 2003, Denver, Colorado
Basin:	Tampico
Countries:	Mexico
Subjects:	Reservoir characterization(45%), Hydrocarbon Systems and Basin Analysis(36%), Petrophysics(34%)

Summary

- There is a huge wealth of information stored as unstructured text
- The discovery of new play types and exploitation technologies has given new value to previously unheeded reports and data sets
- Data managers need to provide the best ways of navigating through increasingly noisy unstructured data collections … and to help clients reflect on the importance of improving the signal to noise ratio of their text content
- A combination of structured master data links, peer reviewed taxonomies and subject index thesauruses, E&P aware geographic tags, and appropriately applied natural language search is likely to provide the best outcome.

Presenter: Scott Tidemann, Vice President Petrosys Middle East & Asia Pacific

Email: scott@petrosys.com.au

PETROSYS digital earth THE NEW SEARCH FOR ENERGY