Efficient Access to Relevant Knowledge Extracted from Geoscience Literature Dedicated to Petroleum Basin Exploration by Using IBM Watson*

X. Guichet[1], N. Dubos-Sallée[1], M.-C. Cacas-Stentz[1], D. Rahon[1], and V. Martinez[2]

Search and Discovery Article #42452 (2019)**
Posted September 23, 2019

*Adapted from oral presentation given at AAPG 2019 Annual Convention & Exhibition, San Antonio, Texas, May 19-22, 2019
 DOI:10.1306/42452Guichet2019

[1]IFPEN (xavier.guichet@ifpen.fr)
[2]TECHadvantage

Abstract

The aim of the study is to enhance the efficiency for collecting relevant geoscientist data among huge amount of unstructured scientific documents by using machine learning algorithm. Valuable knowledge can be found in scientific document collections, however scientists lack of time and are disconcerted to effectively consult mountains of unstructured documents. The main motivation of this work was to create a system able to identify among large repositories what documents are relevant to answer specific questions related to petroleum exploration and more precisely to source rock characterization. The work has been conducted to apply machine learning systems, namely WATSON (IBM) to support geoscientists in a regional geological study. Scientific publications provide information in the form of text, curves or figures. Therefore, two types of machine learning algorithms were tested: one dedicated to image recognition (Watson Visual Recognition WVR) and one to text analysis (Watson Knowledge Studio WKS). First WVR was trained to identify specific image/charts in scientific publications (Event Chart, Stratigraphic column, burial curves and well logs). WVR is able to discriminate efficiently the images of interest from the others even if it was trained with only few dozens of seeds for each image class. Second WKS was trained to understand the semantic framework of textual knowledge related to source rocks. The first step was to list a set of questions we would like to provide answers, e.g. what are the formations bearing source rock in Basin X? What are the Miocene source rock formations in Country X? What is the depositional environment of the source rocks in Basin X? Based on the set of questions and on the recurrence of terms, an ontology (a definition of the entities and relations between entities) was defined. The ontology was willingly limited to ten entities and their relations to make a quick test. WKS has been trained on a set of annotated documents (~150 extracts of ~1000 words). The trained WKS model can identify quite efficiently the entities and the associated relations. Then the two trained models have been applied on a new set of documents, and the extracted information has been stored in a database. The last step was to translate our natural language questions into queries. The result is a list of few documents selected and order with an index of relevance by our system. The proposed workflow is promising thanks to the good performance obtained.

EFFICIENT ACCESS TO RELEVANT KNOWLEDGE EXTRACTED FROM GEOSCIENCE LITERATURE DEDICATED TO PETROLEUM BASIN EXPLORATION BY USING IBM WATSON

X. GUICHET, N. DUBOS-SALLÉE, M.-C. CACAS-STENTZ, D. RAHON *(IFPEN)*

V. MARTINEZ (TECHADVANTAGE)

CONTEXT AND AIMS

- **Context**
  - Deriving additional value thanks to AI from scientific publications: a continuous willingness for decades.
  - To perform regional study, geoscientists have to search information among huge amount of **unstructured documents**.
- IFPEN is running a **feasibility study**
- **Aims**
  - To assess the **ability of Deep Learning** to find the most relevant documents answering to a technical question in a huge amount of documents (scientific publications)
  - To evaluate the **human effort required** to train the model
- Domain of interest: **O&G exploration (source rock characterization)**
- Tested Technology : **WATSON from IBM**

WORKFLOW : 3 MAIN STEPS

NEW ENERGIES

STEP # 1

Preparation of documents

Text and images extractions

Identification of Watson modules

Knowledge Studio

Détails

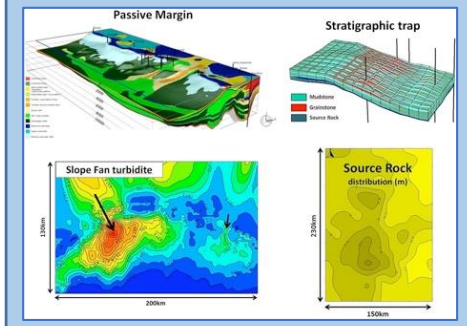
Teach Watson the language of your domain

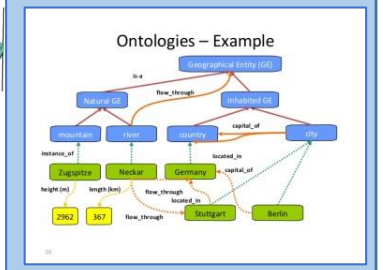
Visual Recognition

Quickly and accurately tag, classify and train visual content

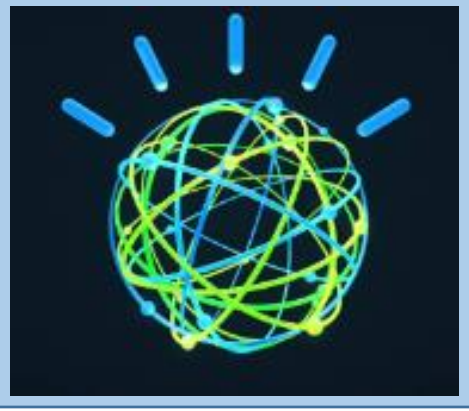
STEP # 2 : Model training

Definition of classes of images / Ontology





Trained Model



STEP # 3
Model deployment

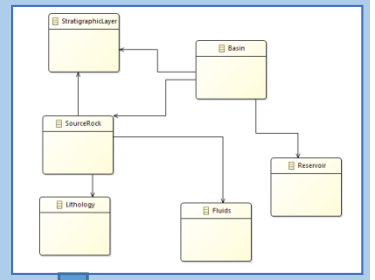
Discovery

Détails Démonstration

Unlock hidden value in data to find answers

Lite IBM

Knowledge graph



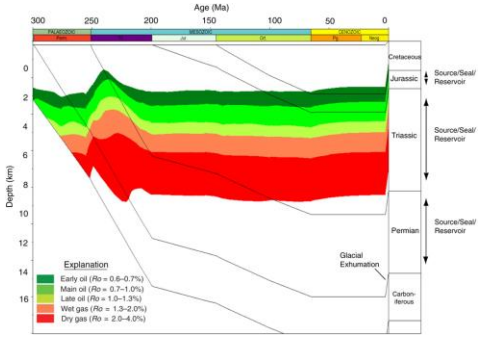
Request

ifp Energies nouvelles

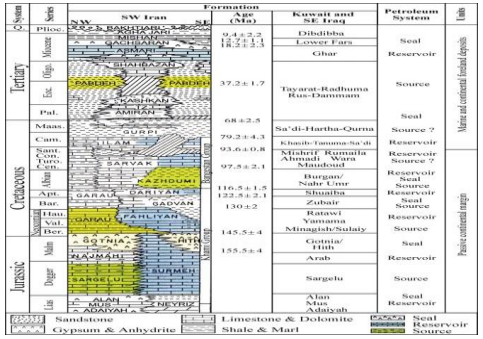
STEP #2
IMAGE CLASSIFICATION (1/2)

NEW ENERGIES

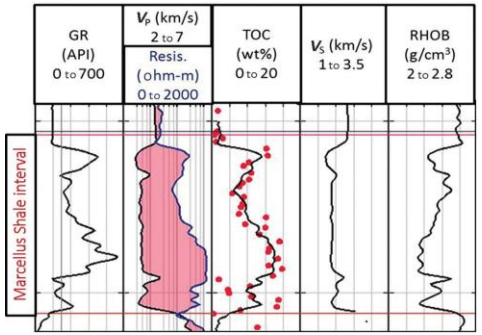
- We trained Watson Visual Recognition on 6 types of graphs with only **few dozens of seeds for each**



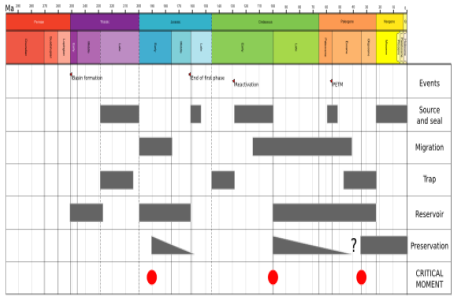
Burial curve



Column stratigraphic

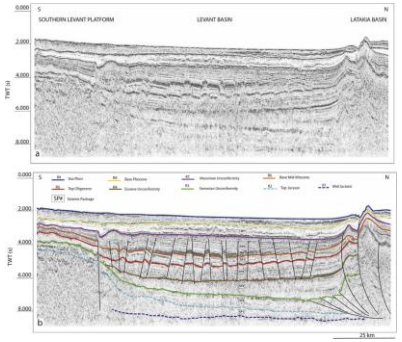


Logging curve

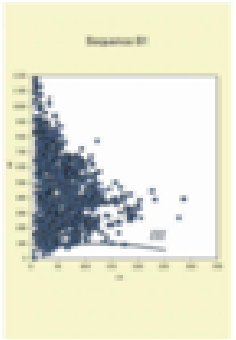


Petroleum system chart

« Trash Classes »



Sections seismic

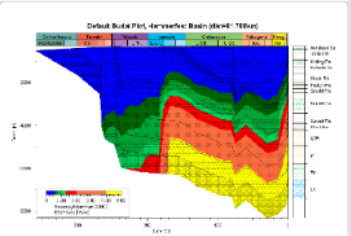
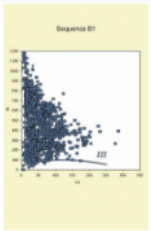
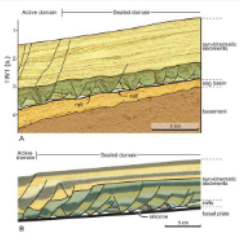
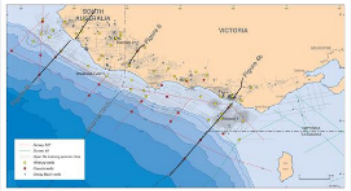
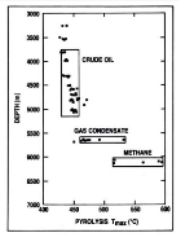


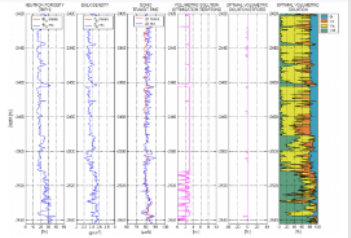
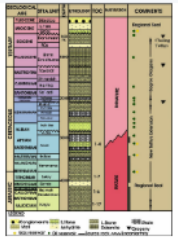
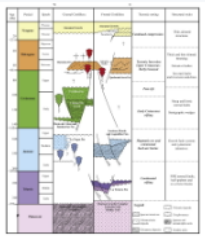
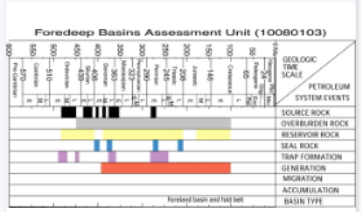
Curves

STEP #2
IMAGE CLASSIFICATION (2/2)

NEW ENERGIES

× Clear results

Burial-history-graph-of-the-Hammerfest-Ba...		m_00001_psisdghhs25_bedretal_page_17_...		m_00017_psisdghhs25_fortbrun_page_16_...		m_00038_psisdghhs25_ryanetal_page_26_...		m_00039_psisdghhs25_sassetal_page_22_...	
									
burial_curve_training	0.91	Curves	0.92	Sections_Sismique	0.92	Sections_Sismique	0.92	Curves	0.92
Sections_Sismique	0.02	Sections_Sismique	0.00	burial_curve_training	0.00	burial_curve_training	0.00	Petroleum_chart_training	0.00
Curves	0.00	Petroleum_chart_training	0.00	Curves	0.00	Curves	0.00	Sections_Sismique	0.00
Petroleum_chart_training	0.00	burial_curve_training	0.00	Petroleum_chart_training	0.00	Petroleum_chart_training	0.00	burial_curve_training	0.00
colonne_strati_training	0.00	colonne_strati_training	0.00	colonne_strati_training	0.00	colonne_strati_training	0.00	colonne_strati_training	0.00
well_logging_training	0.00	well_logging_training	0.00	well_logging_training	0.00	well_logging_training	0.00	well_logging_training	0.00

Optimal-interpretation-of-a-neutron-density...		Stratigraphic-column-of-South-Iraq-Basrah-...		1-s2.0-S0040195115006472-gr3.jpg		F9.large.jpg	
							
well_logging_training	0.92	colonne_strati_training	0.83	colonne_strati_training	0.91	Petroleum_chart_training	0.91
Curves	0.00	well_logging_training	0.36	Curves	0.00	Curves	0.00
Petroleum_chart_training	0.00	Curves	0.00	well_logging_training	0.00	colonne_strati_training	0.00
Sections_Sismique	0.00	Petroleum_chart_training	0.00	Petroleum_chart_training	0.00	well_logging_training	0.00
burial_curve_training	0.00	Sections_Sismique	0.00	burial_curve_training	0.00	Sections_Sismique	0.00
colonne_strati_training	0.00	burial_curve_training	0.00	Sections_Sismique	0.00	burial_curve_training	0.00

Trash classes helped us to better recognize our targets

STEP #2
UNLOCK INFORMATION FROM SCIENTIFIC PUBLICATIONS

- List the **scientific questions** to define the ontology

  Is there a source rock found in the XXX Basin ?

  What is the depositional environment of the XXX formation ?

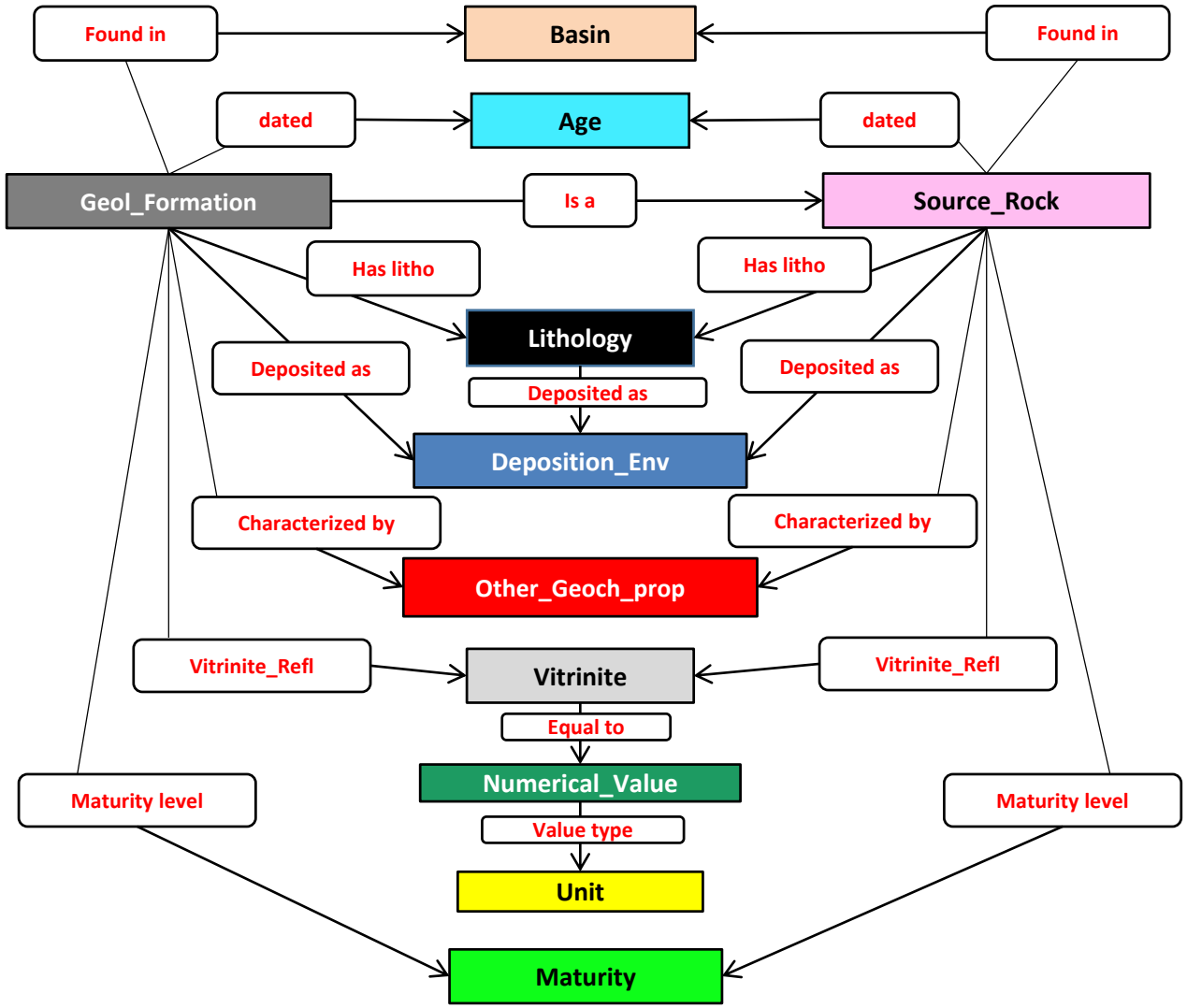
  What are the Miocene formations in Country XXX ?

  What are the formations bearing source rocks in XXX Basins ?

STEP #2
UNLOCK INFORMATION FROM SCIENTIFIC PUBLICATIONS

NEW ENERGIES

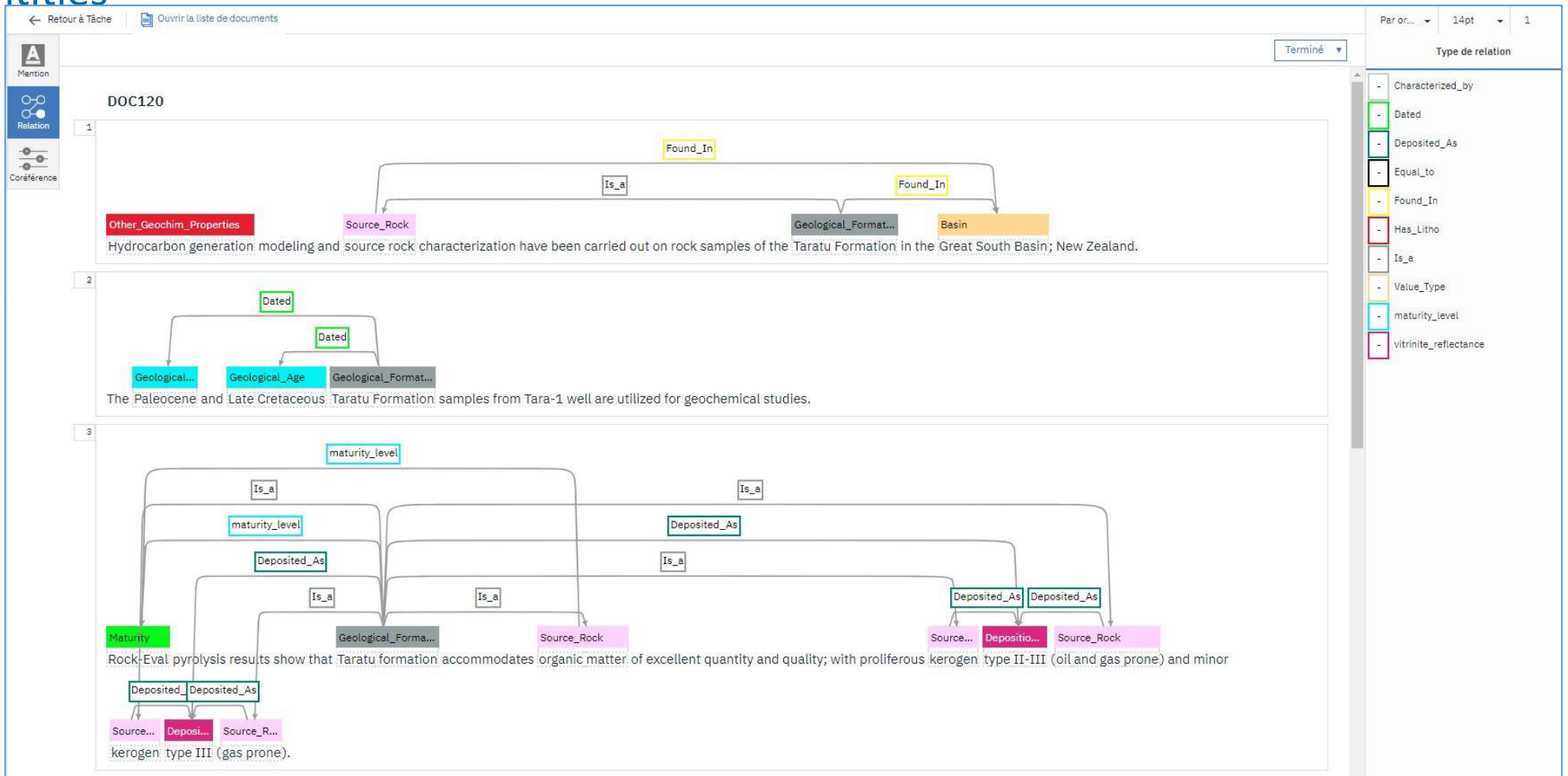
1. List the **scientific questions** to define the ontology
2. Limit the ontology to perform a **quick test**
3. Take into account **short-cuts** made by the authors
4. Select the paper **extracts** for annotation
5. Annotate



STEP #2
UNLOCK INFORMATION FROM SCIENTIFIC PUBLICATIONS

NEW ENERGIES

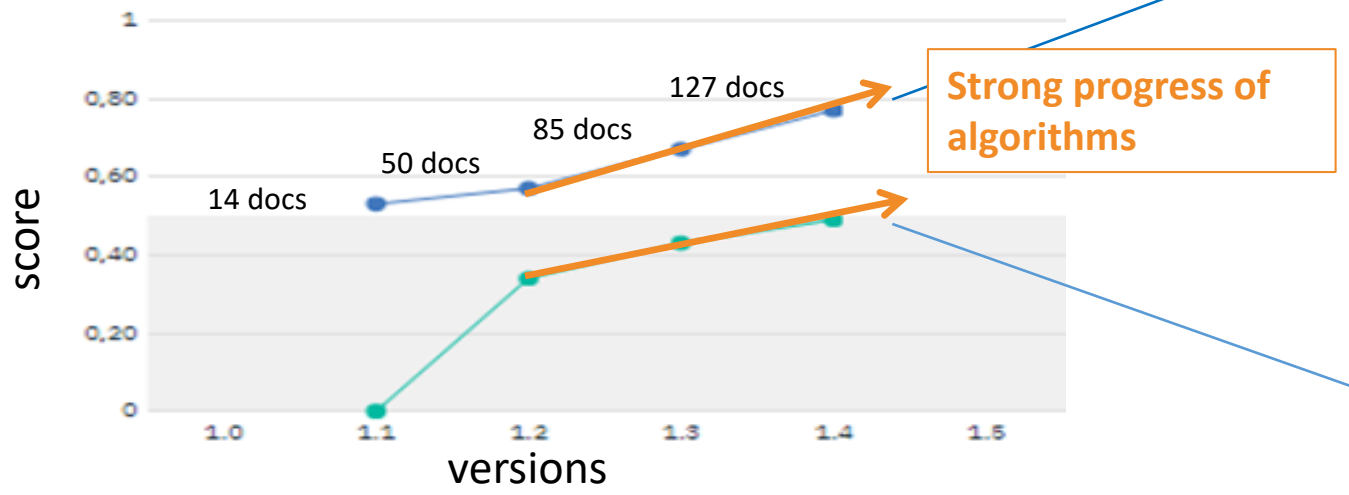
- Annotating extracts using **entities** to teach the system: the key points
- Annotating extracts using **relations** to teach the system the key relations between entities



STEP #2
GLOBAL PERFORMANCE OF THE TRAINED MODEL

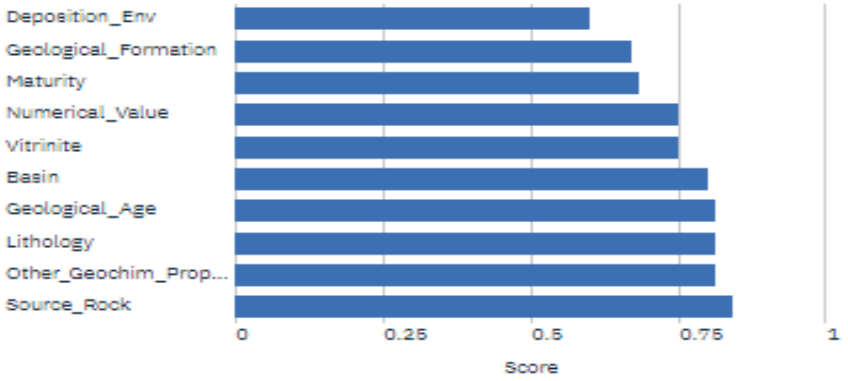
NEW ENERGIES

- Training set: 90% of the annotated docs
- Score = (Precision+Recall)/2



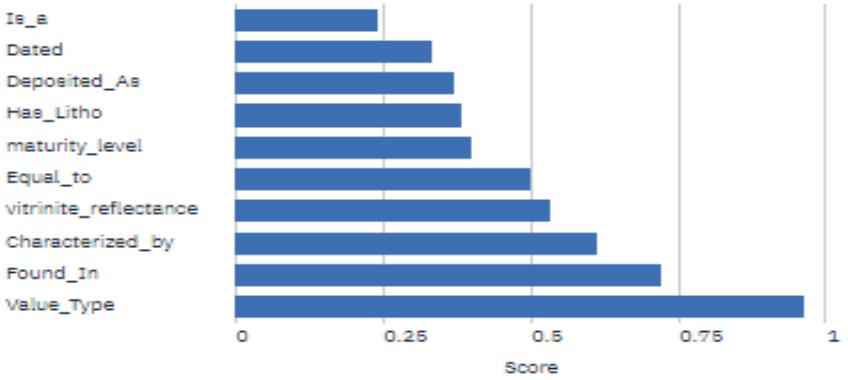
Entities

10 weakest scores



Relations

10 weakest scores



- Good recovery of entities even if the nb of annotated documents is small
- To improve relations score, we have to annotate more documents (>200)

STEP #3 : DEPLOYMENT AND QUERY

NEW ENERGIES

DATA BASE
Of documents

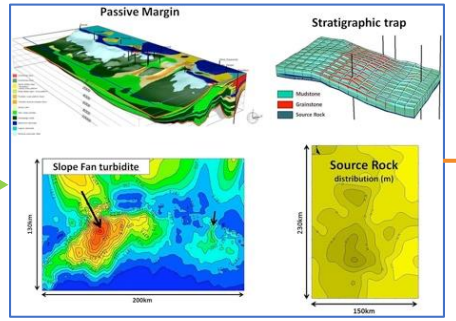
REQUESTS

Answers

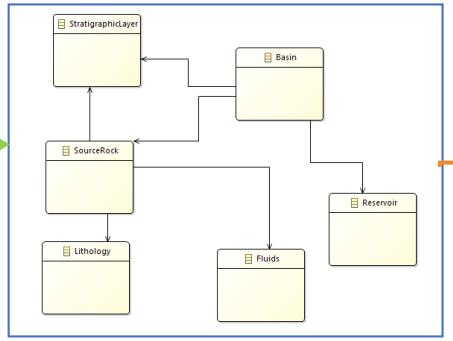
TRAINED MODEL

DEPLOYMENT

Classified images



Graph of knowledge



STEP #3
QUERYING WITH DISCOVERY AND MONGO DB

NEW ENERGIES

- Knowledge graphs generated by Watson Discovery
  - Natural language query ?
    - ➔ **Too difficult**: Training of the system based on the answers relevance is needed
  - Translate the questions into specific IT queries ?
    - ➔ **Very difficult** to build complex queries such as « What are the Miocene formations in Egypt ? » which combine several entities and relations.

- Export the Knowledge Base into MongoDB and exploit the Base using Studio 3T
  - Some questions are translated into IT queries
  - Easier than Discovery but not straightforward.

Studio 3T

- Clearly the most difficult step :
  - Exploitation of image classification and text understanding… in progress !

CONCLUSIONS & PERSPECTIVES

NEW ENERGIES

- Image Classifier using Visual Recognition works very well and very quickly (2-3 man-days)
  - *Next Step :* Exploitation of the information contained in the classified images (like symbols)
- Text model built with Knowledge Studio already works well with only 127 annotated documents and only ~15 man-days.
  - *Next Step :* Improvement of the training, particularly for relations (only time consuming)
- Trained model shows good performances : our workflow is promising
  - *Next Step :* Better combine text and image information in the knowledge graph
- Query phase is finally far more tricky than expected
  - Need to improve the query phase to exploit the full capability of the workflow
- Next Step : Work with partners curious about this test

Innovating for energy

Find us on:

www.ifpenergiesnouvelles.com

@IFPENinnovation