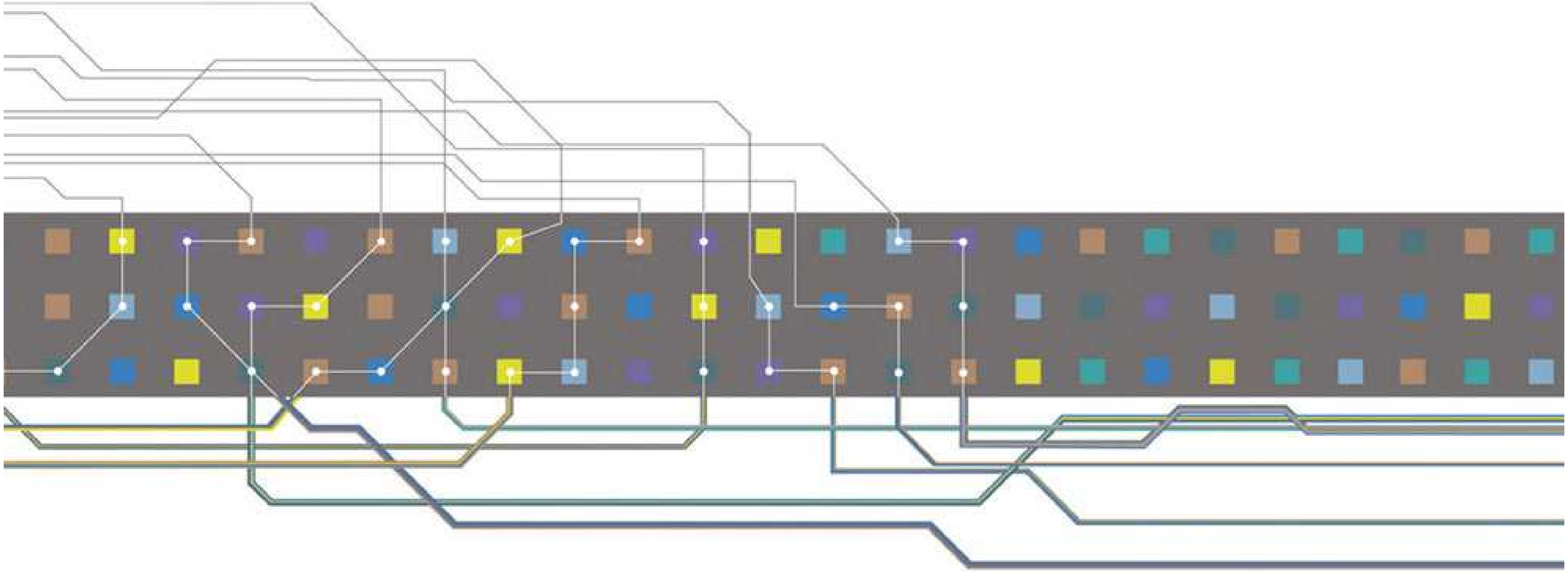
THALES



Technological Challenges for Multimodal Information Mastering

Workshop Franco-Finlandais 5/5/06

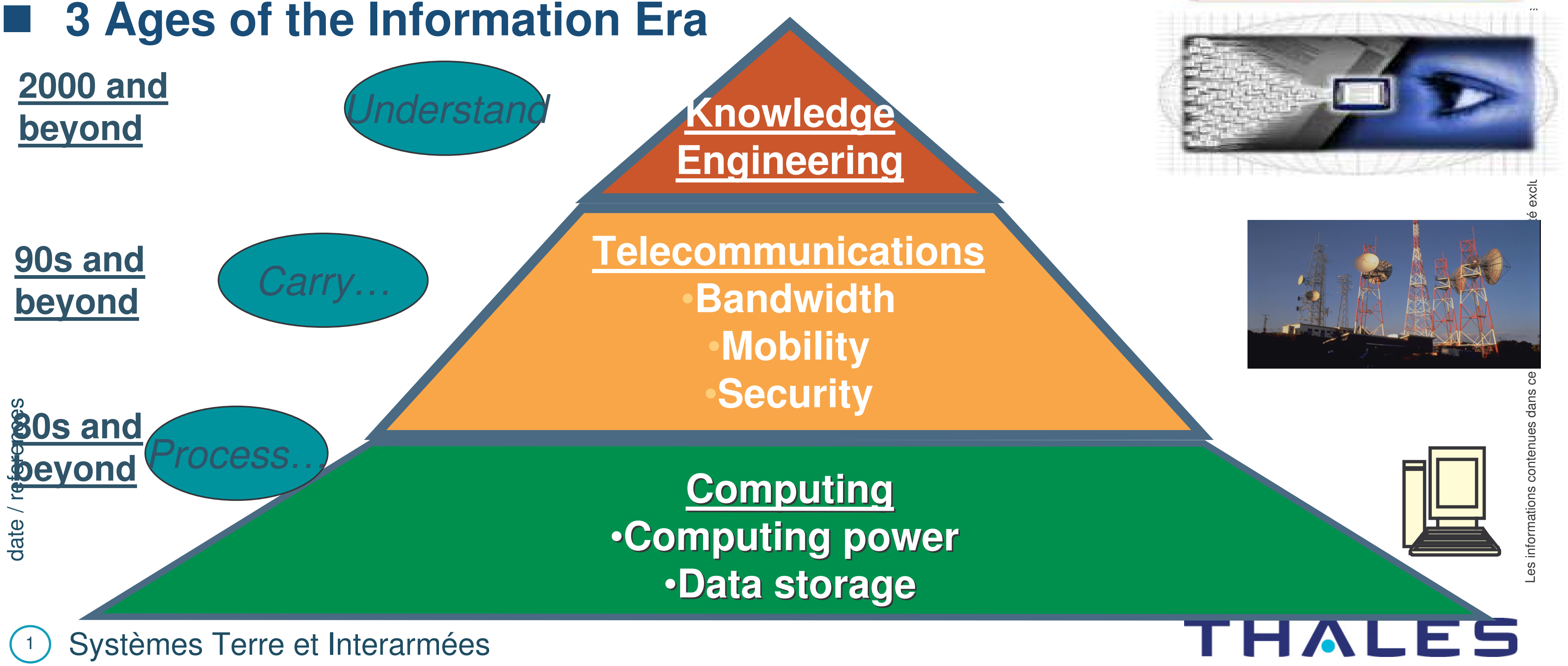
François Marcotorchino

date / references

Systèmes Terre et Interarmées

New Trends

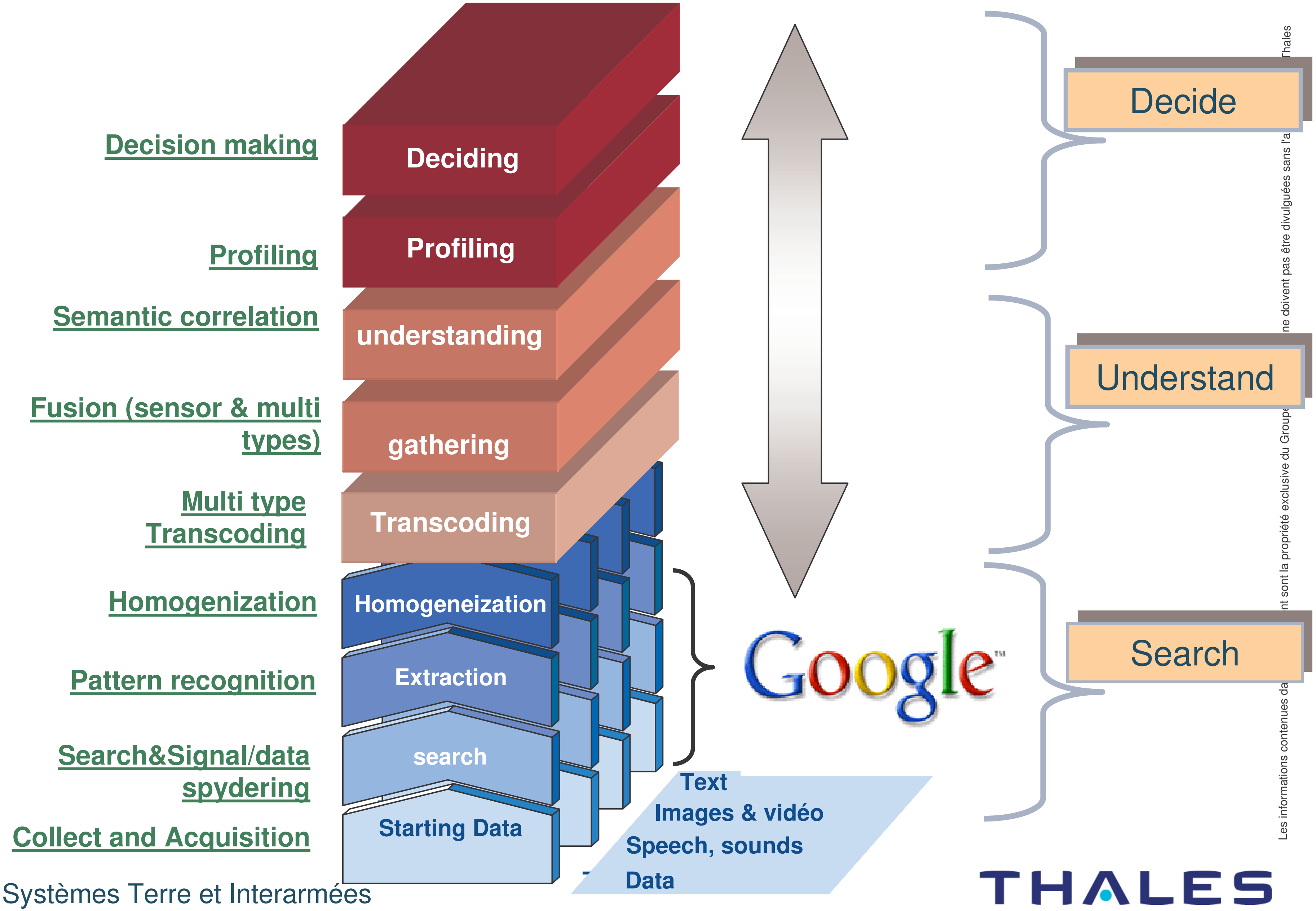
- Data Growth
  - Quantity of available data worldwide doubles every year
- Availability of processing tools
  - Cost of data storage has decreased from €/Mb to cents / Mb
  - Computing power continues to double every year
- 3 Ages of the Information Era



Technological Challenges of Multimodal Information Mastering

- Be at the level of the best for the indexation of the web
- Improve the current processes of consultation
- Drastically improve "Miners*" technology
- Merge and inter-correlate multi-source data
- Achieve breakthrough in « machine automata » technology to improve *extraction* quality
- Optimize the transcoding process
- Achieve true interoperability between technologies
- Manage multimedia bases with the same efficiency as current text or digital data databases
- Develop applications in the Domains of : Security, Health, E-training, Digital libraries, Finance, and Digital Life

Knowledge Tower



Decision making

Profiling

Semantic correlation

Fusion (sensor & multi types)

Multi type Transcoding

Homogenization

Pattern recognition

Search&Signal/data spydering

Collect and Acquisition

THALES

Technological Keywords for Providing the systems with Advanced Information Analysis

- Information Extraction
- Automatic Clustering (Structured/unstructured)
- Transcoding Data from Type A to Type B
- Information Fusion
- Semantic Graphs

High level of Difficulties but Challenging Potential Improvements

Different Types of Data Fusion Type A➔Type B

- **Fusion: Text ➔Numerical Structured Data**

  This fusion (or joint treatment Textual Data with Structured Data) must be understood as semantic reciprocal treatments, allowing semantic correlation and not "simple juxtaposition"

- **Fusion : Speech ➔ Text** (already under process)

  Transformation by means of transcoders of the "speech signal" into its "textual translation" (example : Voice Dictation IBM, LIMSI tools etc..)

- **Fusion : Image➔ Structured Numerical Data** (more classical)

  Here we must use tools for "image contours and outlines analysis" for translate Image into "Semantic Descriptive Qualitative Markers"

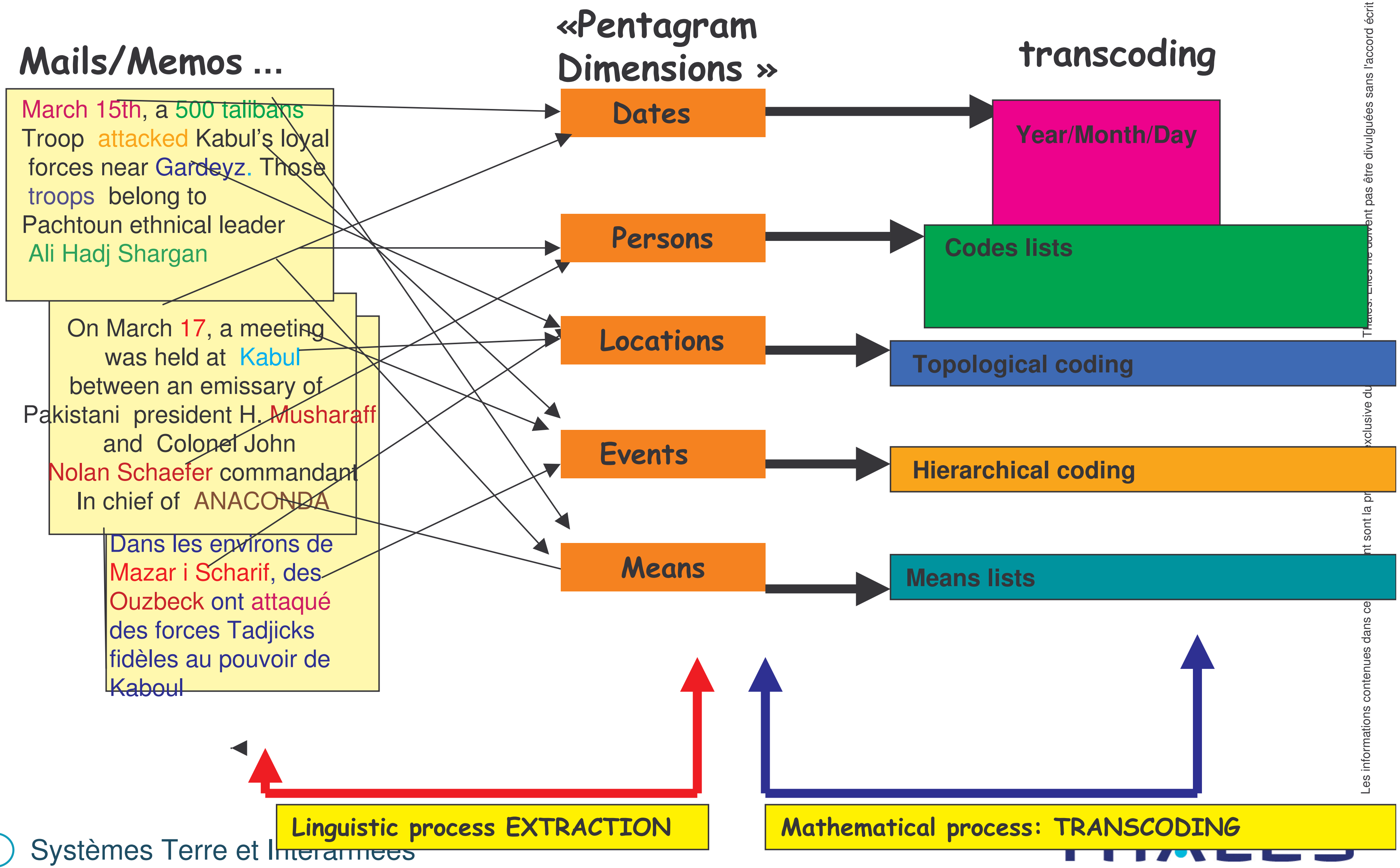
- **Fusion : Image ➔ Text**

  Here we must try to interpret (if possible) the contours and forms by means of predetermined structures (gabarits) with associated generation of textual scripts.

- **Fusion: Cartographical Data ➔ Structured Numerical Data**

  Dans In that case we translate the topographical Information into a set of Qualitative (hierarchical or Nominal) compatible with the other descriptive numerical variables already stored.

Extracted information transcoding process: Text ➔ Data



Mails/Memos ...

March 15th, a 500 talibans Troop attacked Kabul's loyal forces near Gardeyz. Those troops belong to Pachtoun ethnical leader Ali Hadj Shargan

On March 17, a meeting was held at Kabul between an emissary of Pakistani president H. Musharaff and Colonel John Nolan Schaefer commandant In chief of ANACONDA

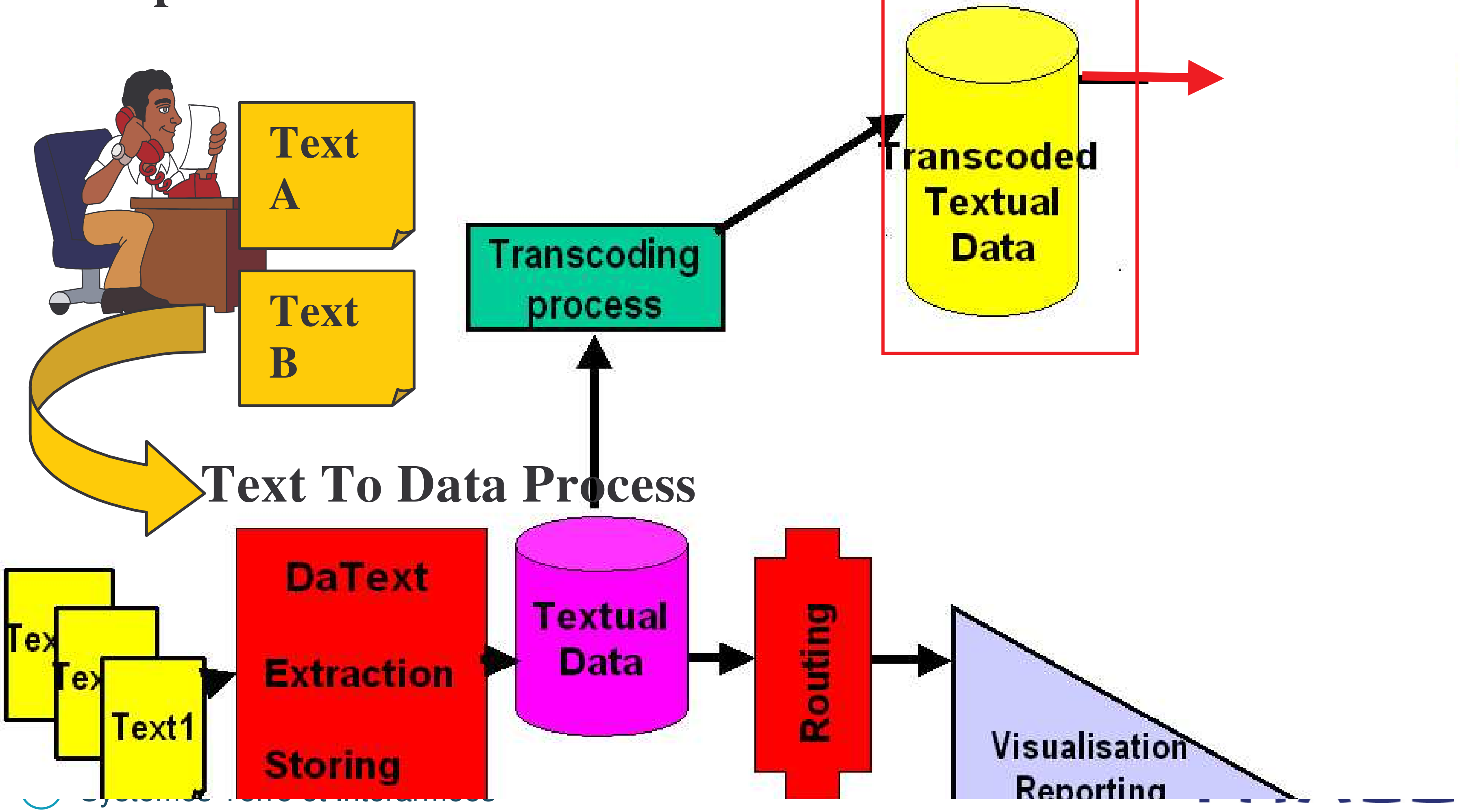
Dans les environs de Mazar i Scharif, des Ouzbeck ont attaqué des forces Tadjicks fidèles au pouvoir de Kaboul

Linguistic process EXTRACTION

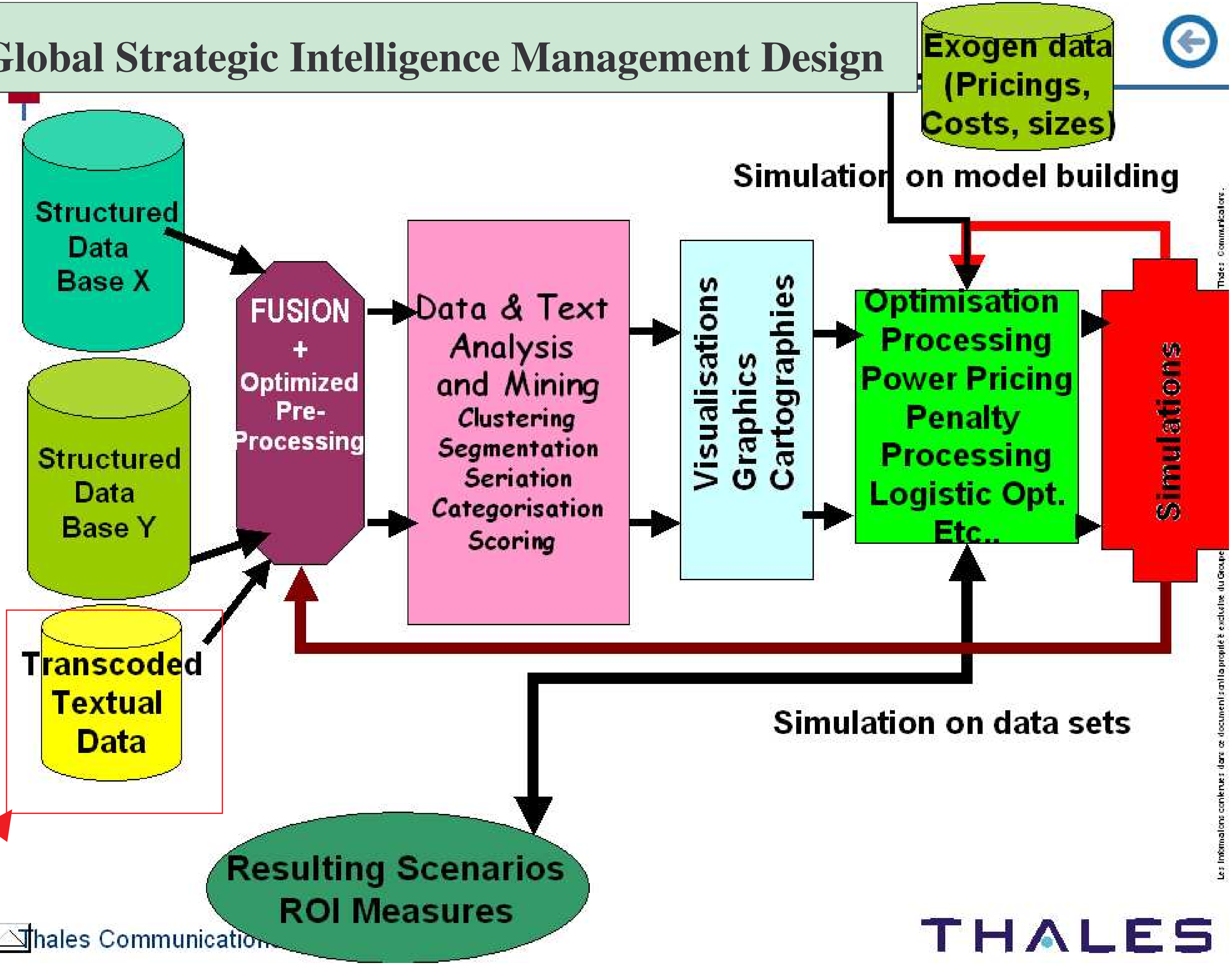
Mathematical process: TRANSCODING

Speech to text and Information Extraction Chain Coupling

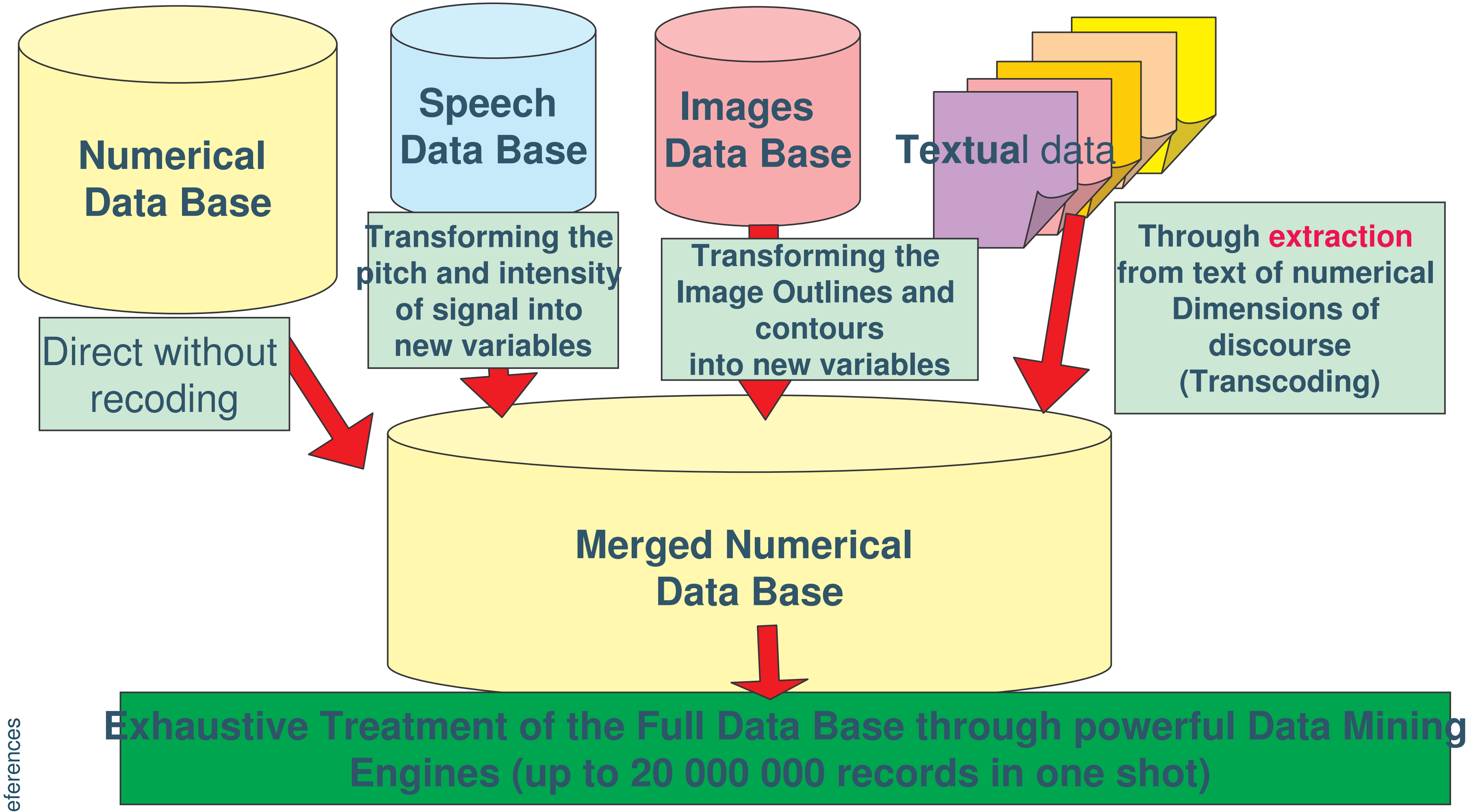
Speech to text Process



Global Strategic Intelligence Management Design



Semantic Fusion of Multimodal Information (Case1)

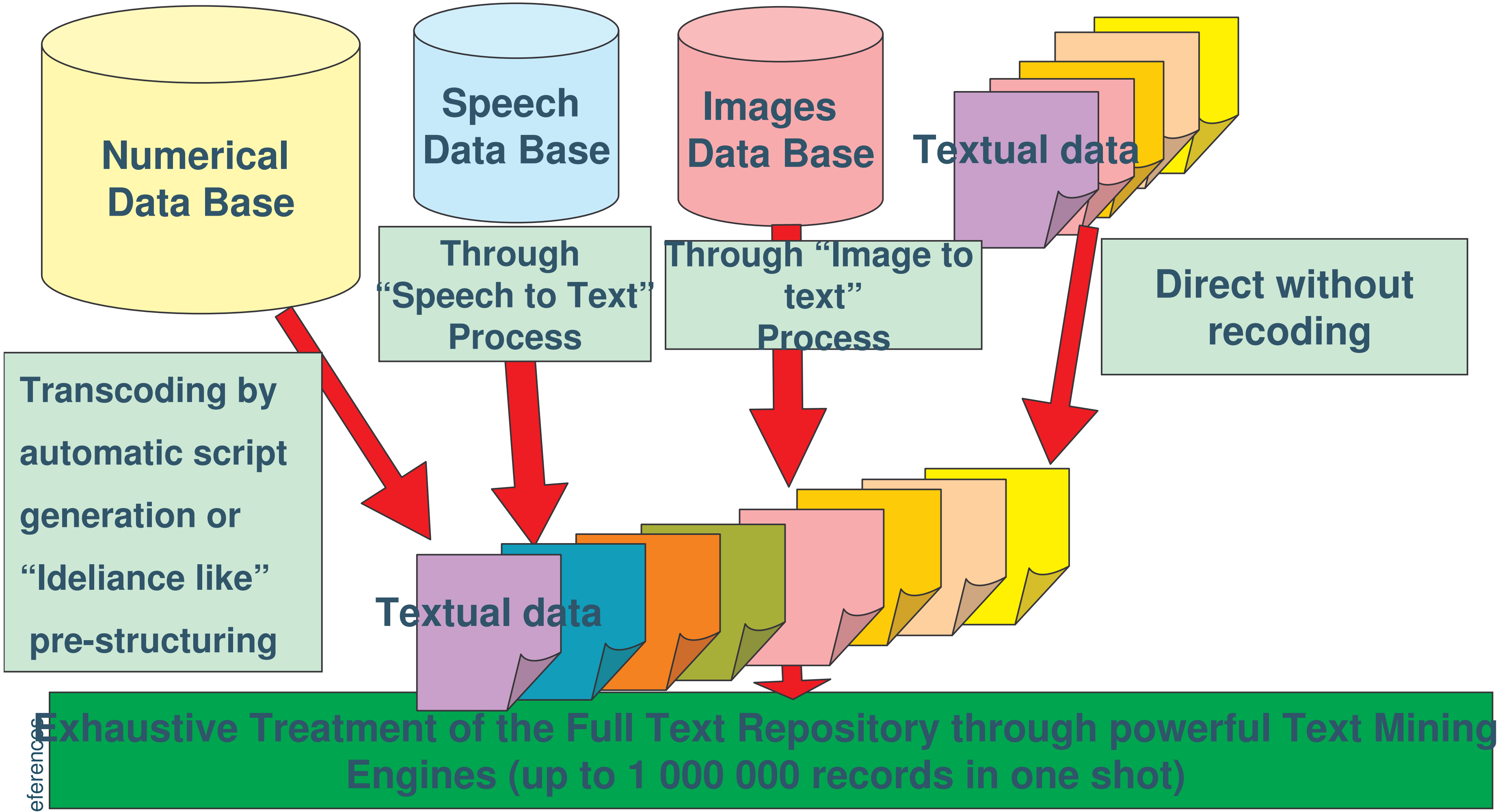


Numerical Data Base
Speech Data Base
Images Data Base
Textual data
Direct without recoding
Transforming the pitch and intensity of signal into new variables
Transforming the Image Outlines and contours into new variables
Through extraction from text of numerical Dimensions of discourse (Transcoding)
Merged Numerical Data Base
Exhaustive Treatment of the Full Data Base through powerful Data Mining Engines (up to 20 000 000 records in one shot)

First Strategy: Transform all the set of data into structured ones:

Very powerful for huge sized treatments but loss of semantical value

Semantic Fusion of Multimodal Information (Case2)



Numerical Data Base

Speech Data Base

Images Data Base

Textual data

Through "Speech to Text" Process

Through "Image to text" Process

Direct without recoding

Transcoding by automatic script generation or "Ideliance like" pre-structuring

Textual data

Exhaustive Treatment of the Full Text Repository through powerful Text Mining Engines (up to 1 000 000 records in one shot)

Second Strategy: Transform the whole data set into textual information : Very powerful for keeping a high level of semantic value, less powerful for large sized problems

The trilogy: Semantic Network representation /Analysis / Clustering

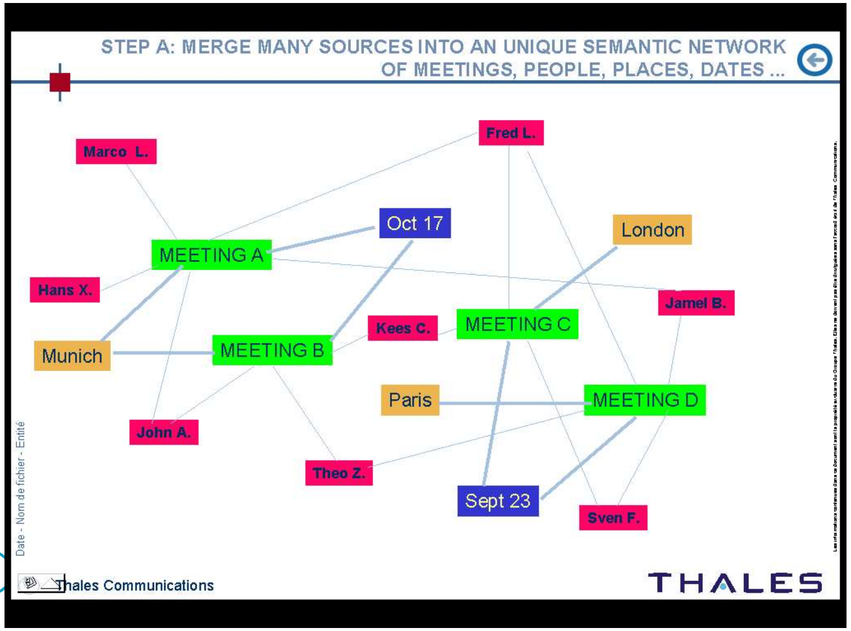
From a SEMANTIC GRAPH, a large variety of Problems related to Clustering and Similarity is appearing:

A each node of the Graph one gives the attributes which are functions of the linked sub-graphs

Consider   those sub-graphs per say -as new objects to classify (or to cluster)

Research all the chains of the graphs which constitute the attributes for their classification or clustering

The Clustering results are a real enrichment for the Semantic Graph and go directly into the *TURBO* machine for other treatments.



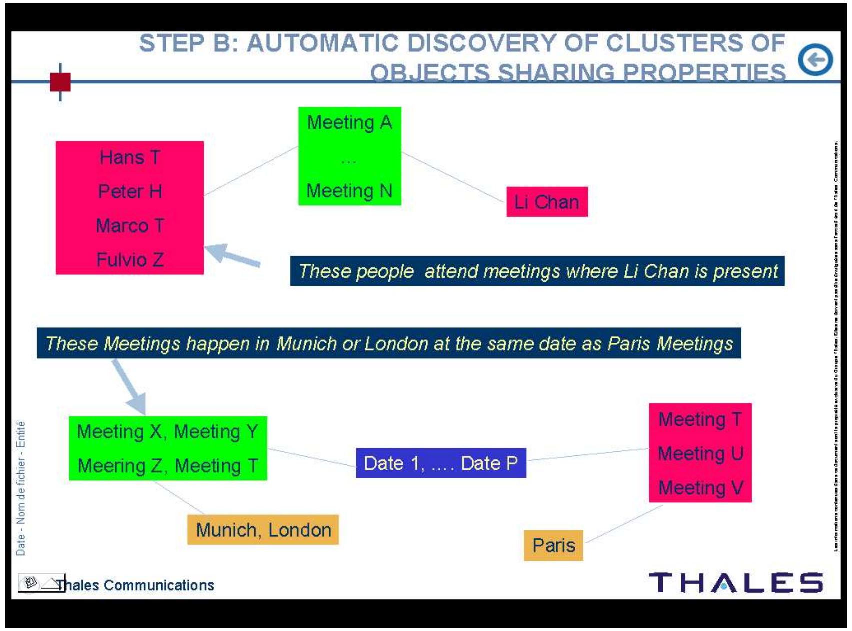
STEP A: MERGE MANY SOURCES INTO AN UNIQUE SEMANTIC NETWORK OF MEETINGS, PEOPLE, PLACES, DATES ...

Marco L.
Fred L.
Oct 17
London
MEETING A
Hans X.
Jamel B.
Kees C.
MEETING C
Munich
MEETING B
Paris
MEETING D
John A.
Theo Z.
Sept 23
Sven F.

Date - Nom de fichier - Entité

Thales Communications

THALES



STEP B: AUTOMATIC DISCOVERY OF CLUSTERS OF OBJECTS SHARING PROPERTIES

Meeting A
...
Meeting N

Hans T
Peter H
Marco T
Fulvio Z

Li Chan

These people attend meetings where Li Chan is present

These Meetings happen in Munich or London at the same date as Paris Meetings

Meeting X, Meeting Y
Meering Z, Meeting T

Date 1, ... Date P

Meeting T
Meeting U
Meeting V

Munich, London

Paris

Date - Nom de fichier - Entité

Thales Communications

THALES

Search Industry Facing Evolution

What will Google do next? How will Yahoo counter the move? And how will this impact MSN Search, Ask Jeeves and the scores of smaller players vying to get their share of the search engine revenue pie?

Clustering Catches On

*« Another common prediction is for **the increased adoption of clustering technology**. America Online is already offering clustering via its Vivisimo partnership. **Gartner analyst Alan Weiner**, for one, said he's hooked on clustering, and he **expects to see major search engine players add clustering features in 2005** to make search more user-friendly.*

Clustering is an elemental way of taking people through a more direct path to what they are looking for," Gartner analyst said. "If you type in the world 'polish,' the search engine might not know if you are looking for information about Poland or products that make your car shiny. With clustering technology, you have on-the-fly categories and you can immediately choose 'car-related accessories. »

Search Industry Evolution towards Clustering

Enterprise Search summit

May 23-24, 2006
Hilton New York, New York City

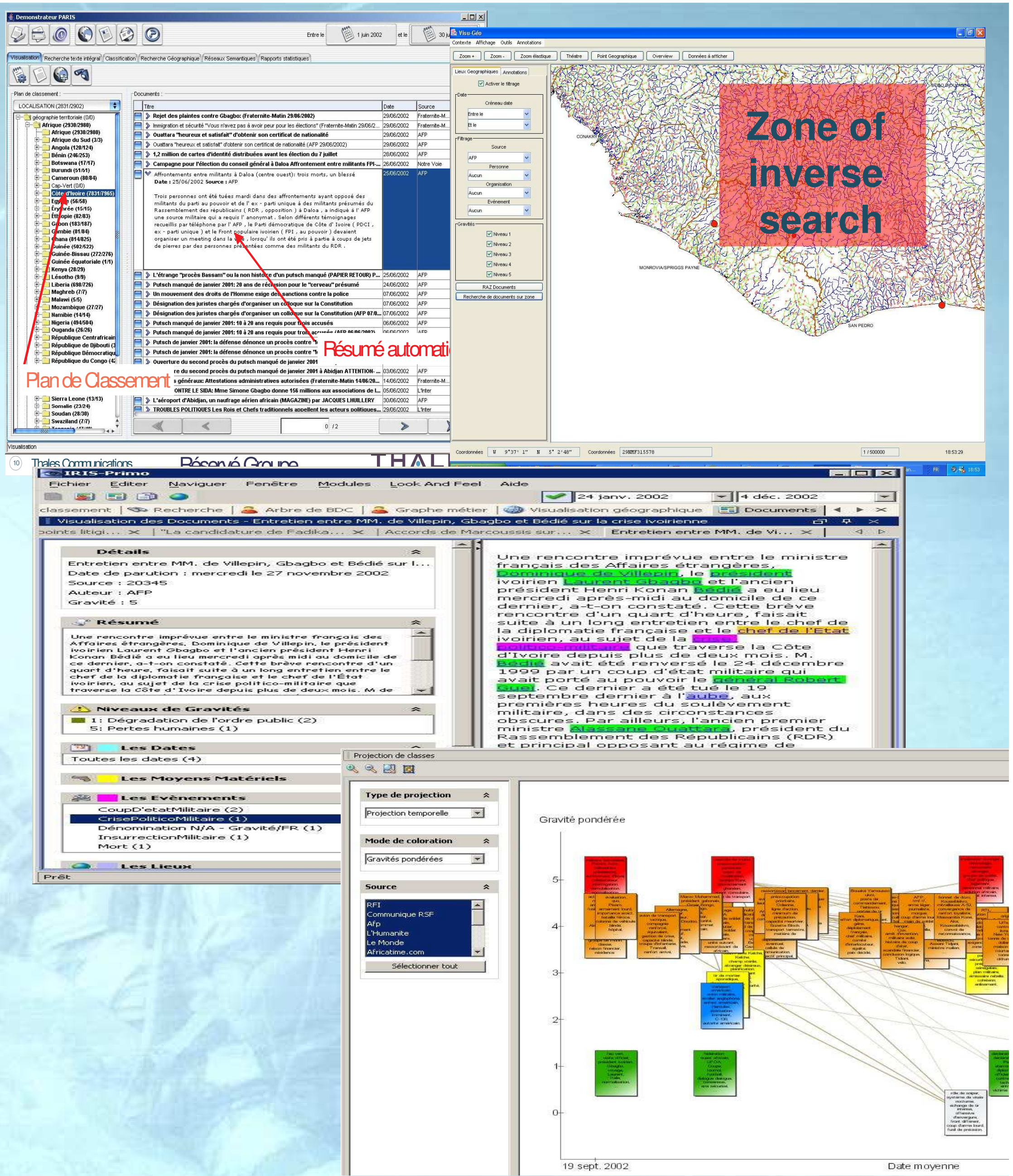
Beyond Search: Intelligent Use of Intelligence

William Lunceford, Section Manager, Procter & Gamble

« Search tools are more abundant than ever, but the people you support are spending valuable time sifting through irrelevant search results and may be missing critical information. How can you improve their overall search experience, and, more importantly, how can you help them make intelligent use of search intelligence? P&G created a unified search tool that sorts results **into clusters** that are intelligently selected from words and phrases found in the documents themselves. Learn about the benefits of clustering and how P&G's project evolved to extend across the entire enterprise ». *Hosted by Vivisimo.*

« Sprint Platform™ » functionnal spectrum

- **Automatic treatments of Documents**
  - Indexation/categorization/classification plans (**pentagramm**)
  - Multi-modal Filtering, highlighting of « named entities »
  - Hierarchy of Documents according to « **gravity or criticality** » levels
  - Automatic Summaries
  - Reporting
- **Dynamic Cartography**
  - Automatic Geolocation of events
  - Geographical **Inverse search** of documents
  - Filtering and annotations
- **Decision Aided process**
  - **Clustering**
  - Semantic Networks



Détails

Entretien entre MM. de Villepin, Gbagbo et Bédié sur l...
Date de parution : mercredi le 27 novembre 2002
Source : 20345
Auteur : AFP
Gravité : 5

Résumé

Une rencontre imprévue entre le ministre français des Affaires étrangères, Dominique de Villepin, le président ivoirien Laurent Gbagbo et l'ancien président Henri Konan Bédié a eu lieu mercredi après midi au domicile de ce dernier, a-t-on constaté. Cette brève rencontre d'un quart d'heure, faisait suite à un long entretien entre le chef de la diplomatie française et le chef de l'État ivoirien, au sujet de la crise politico-militaire que traverse la Côte d'Ivoire depuis plus de deux mois. M de

Niveaux de Gravités

1: Dégradation de l'ordre public (2)
5: Pertes humaines (1)

Les Dates

Toutes les dates (4)

Les Moyens Matériels

Les Evènements

CoupD'etatMilitaire (2)
CrisePoliticoMilitaire (1)
Dénomination N/A - Gravité/FR (1)
InsurrectionMilitaire (1)
Mort (1)

Les Lieux

Une rencontre imprévue entre le ministre français des Affaires étrangères, Dominique de Villepin, le président ivoirien Laurent Gbagbo et l'ancien président Henri Konan Bédié a eu lieu mercredi après-midi au domicile de ce dernier, a-t-on constaté. Cette brève rencontre d'un quart d'heure, faisait suite à un long entretien entre le chef de la diplomatie française et le chef de l'Etat ivoirien, au sujet de la crise politico-militaire que traverse la Côte d'Ivoire depuis plus de deux mois. M. Bédié avait été renversé le 24 décembre 1999 par un coup d'état militaire qui avait porté au pouvoir le général Robert Gueï. Ce dernier a été tué le 19 septembre dernier à l'aube, aux premières heures du soulèvement militaire, dans des circonstances obscures. Par ailleurs, l'ancien premier ministre Alassane Ouattara, président du Rassemblement des Républicains (RDR) et principal opposant au régime de

Type de projection: Projection temporelle

Mode de coloration: Gravités pondérées

Source

RFI
Communique RSF
Afp
L'Humanite
Le Monde
Africatime.com

Sélectionner tout

An Example of a French National Initiative for Building up a global Information Search, Extraction and Analysis Solution

INFOM@GIC

Project Objectives

- *Knowledge Engineering* : Develop worldclass technologies enabling « Ile de France » companies to gain lead positions on a fast growing market (> 20B€ 2008)
- Capitalize on the unique scientific potential of Ile-de-France companies, labs, universities, start-ups by federating efforts

Enable everyone, private or professional, to *navigate* easily and smartly in huge amount of multi-source information so as to *understand* and *decide*

Partners

INFOM@GIC

PARIS 8 Cognition & Usages

LSTA UNIVERSITE PIERRE & MARIE CU LA SCIENCE A PARIS

EADS

LIPN UNIVERSITÉ PARIS 13

ina

Odile Jacob

INRIA ROCQUENCOURT

CNRS LIMSI

Vecsys

eDF

Hi-Stor Technologies Innovation in Data Storage

SINEQUA

TELECOM PARIS école nationale supérieure des télécommunications

XEROX

ONERA

LIP6

THALES

EMIS

CEREMADE

Gaspard Monge Institut Université de Marne-la-vallée

PERTIMM

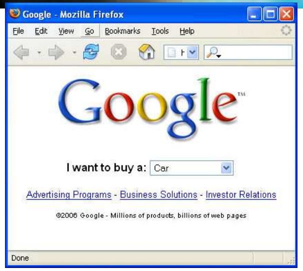
cea

bertin technologies la maîtrise de l'innovation

date / references

INFOM@GIC Multimédia & Vie Numérique

Défis Technologiques



I want to buy a: Car

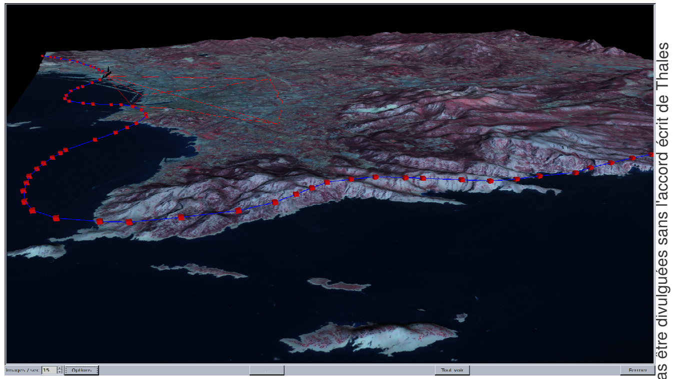
Advertising Programs - Business Solutions - Investor Relations

©2006 Google - Millions of products, billions of web pages

- Etre au niveau des meilleurs pour l'indexation du Web
- Améliorer les processus de consultation actuels
- Avoir de meilleurs « Miners* » qu'IBM ou ORACLE
- Fusionner et inter-corréler les données de nature et sources différentes au delà de l'Etat de l'Art actuel
- Dépasser les limites actuelles des technologies d'automates pour gagner de la performance au niveau Extraction
- Optimiser les transcodages « Type de données X vers Type de données Y »
- Passer d'une simple juxtaposition à une réelle interopérabilité des outils.
- Gérer des bases multimédias avec la même efficacité que la gestion actuelle des bases textuelles ou de données numériques
- Trouver des applications différenciatrices dans les domaines de la Sécurité, de la Santé, de la Finance, et de la Vie Numérique

date / references

Infom@gic facts



- **Budget : 65M€**
- **Duration : 3 years**
- **33 partners**
- **Calendar :**
  - 2006 : models and implementation of first prototypes
  - 2007 : complete suite of prototypes
  - 2008 : technological integration platform and software bricks

Budget split

- Big firms and large Institutes : **58,6%**
- Small and Medium Sized Companies : **21,2%**
- Research Centres and Universities :**20,2%**