The truth of the data pudding is in the integrating

Dr Frédérique Segond

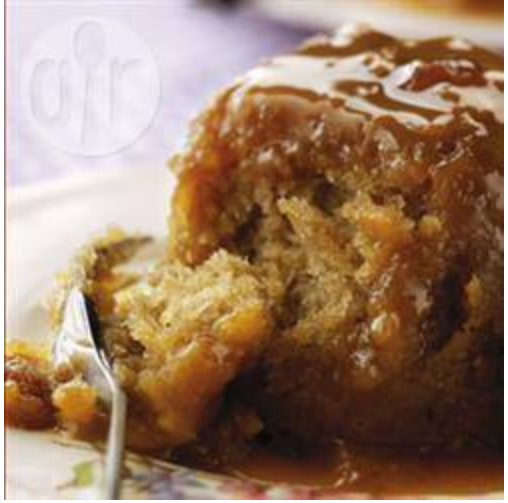
Research Director Viseo,
Associated Professor, INALCO

European Data Forum
November, 16, 17, 2015, Luxembourg

VISEO

www.viseo.com

Summary



Ingredients
Serves: 6
•190g plain flour
•1 1/2 teaspoons baking powder
•120g soft brown sugar
•1 pinch salt
•120ml milk
•2 eggs
•6 tablespoons melted butter or margarine
•2 teaspoons vanilla extract

- **The data pudding : ingredients**
- **Data integration : some of the remaining challenges**
- **The cooking device : the biggest challenge of today**

VISEO

The ingredients of the pudding

Which are the data we want to integrate ?

Data from all over, of any type :

In the Past

Today

How to make sense of the new world ?

Zooming on heterogeneous textual data (remaining challenges) :

- Correct and degradated texts
- Text, numbers, logs
- Structured and unstructured
- BI plateforms efficiency with massive data (scalability)

VISEO

Integrating (analysing) degraded text

SMS :

Lol nan tkt on ta pas oublié
Je suis ds le train, todo bene. Trop bon les fondants au petit dej! Gracias la madre, bonne semaine
Oui cme sa i pouron ri1 te reproché. C complet ce soir?
Oui grav .c mieu.ta vu ia pa 2raisn 2se fair d film .lol

Twitter :

Contre-courants @HKNSCC 10 oct.
C'est en ce moment au #RefletMédicis #Pakeezah #IndianCinema #old #classics
Merci @OSidre pour cette critique de #Pakeezah ! #ContreCourants @HKNSCC

Facebook :

YYY I missed it, but the episode recording should show up here soon :http://www.radio.org/.../www24da.../the-good-man
Can't wait!
XXX: You didn't miss it yet! I'll be on tomorrow. smile emoticon
YYY DOH!
YYY #readstheoriginalpostcarefully

Technical texts :

Vérifier que l'OP EC applique le RDC O fiche N° PCO28 page 13 (Perte RRX voie A)
- Débrochage des cellules des pompes RRI et SEC voies A et B
- Réalimentation des échangeurs RCV 006, 10 et 20EC, CXD et communs par RRV tranche jumelle

Stack over flow :

To do this on my server I had to first: install the URL Rewrtie Modulehttp://www.iis.net/downloads/microsoft/url-rewrite
And then I had to add a web.config file with this XML (this works for removing the .php, if added, as well as adding the .php invisibly to the URL):

```
<?xml version="1.0" encoding="UTF-8"?> <configuration> <system.webServer> <rewrite> <rules>
<rule name="Redirect .php extension" stopProcessing="false"> <match url="^(.*).php$"
ignoreCase="true" /> <conditions logicalGrouping="MatchAny"> <add input="{URL}"
pattern="(.*).php$" ignoreCase="false" /> </conditions> <action type="Redirect" url="{R:1}"
redirectType="Permanent" /> </rule> <rule name="hide .php extension" stopProcessing="true">
<match url="^(.*)$" ignoreCase="true" /> <conditions> <add input="{REQUEST_FILENAME}"
matchType="IsFile" negate="true" /> <add input="{REQUEST_FILENAME}" matchType="IsDirectory"
negate="true" /> <add input="{REQUEST_FILENAME}.php" matchType="IsFile" /> </conditions>
<action type="Rewrite" url="{R:0}.php" /> </rule> </rules> </rewrite> </system.webServer>
</configuration>
```

Integrating degraded text

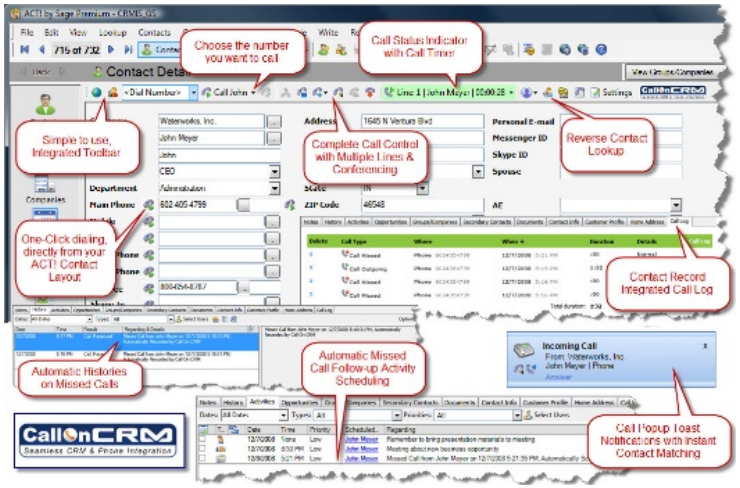
SMS4Sciences

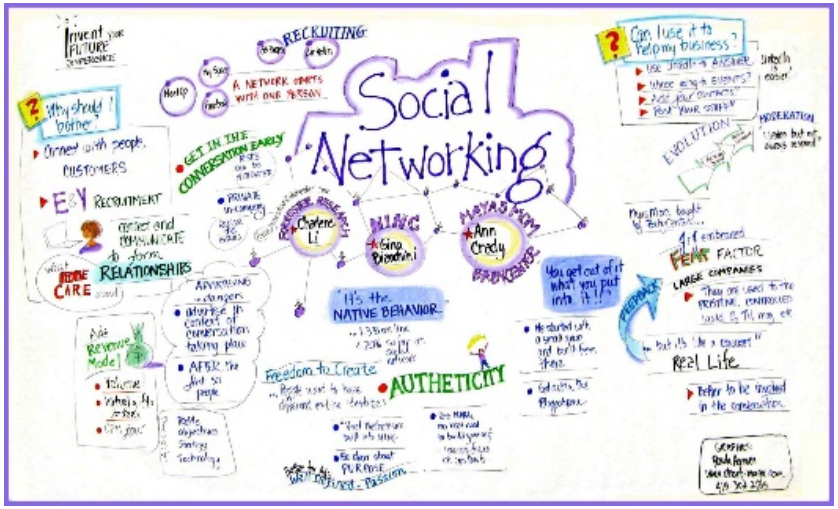
The aim of the sms4scienceproject is to contribute to the study of SMS communication and of its language.
In order to reach this target, researchers from various countries are joining forces to undertake the composition, for a large number of languages, of a vast SMS corpus for scientific research.
15 universities are partners in this project and are carrying out the composition of corpora for their countries.

http://www.sms4science.org
http://sud4science.org/

VISEO

Integrating structured and unstructured data





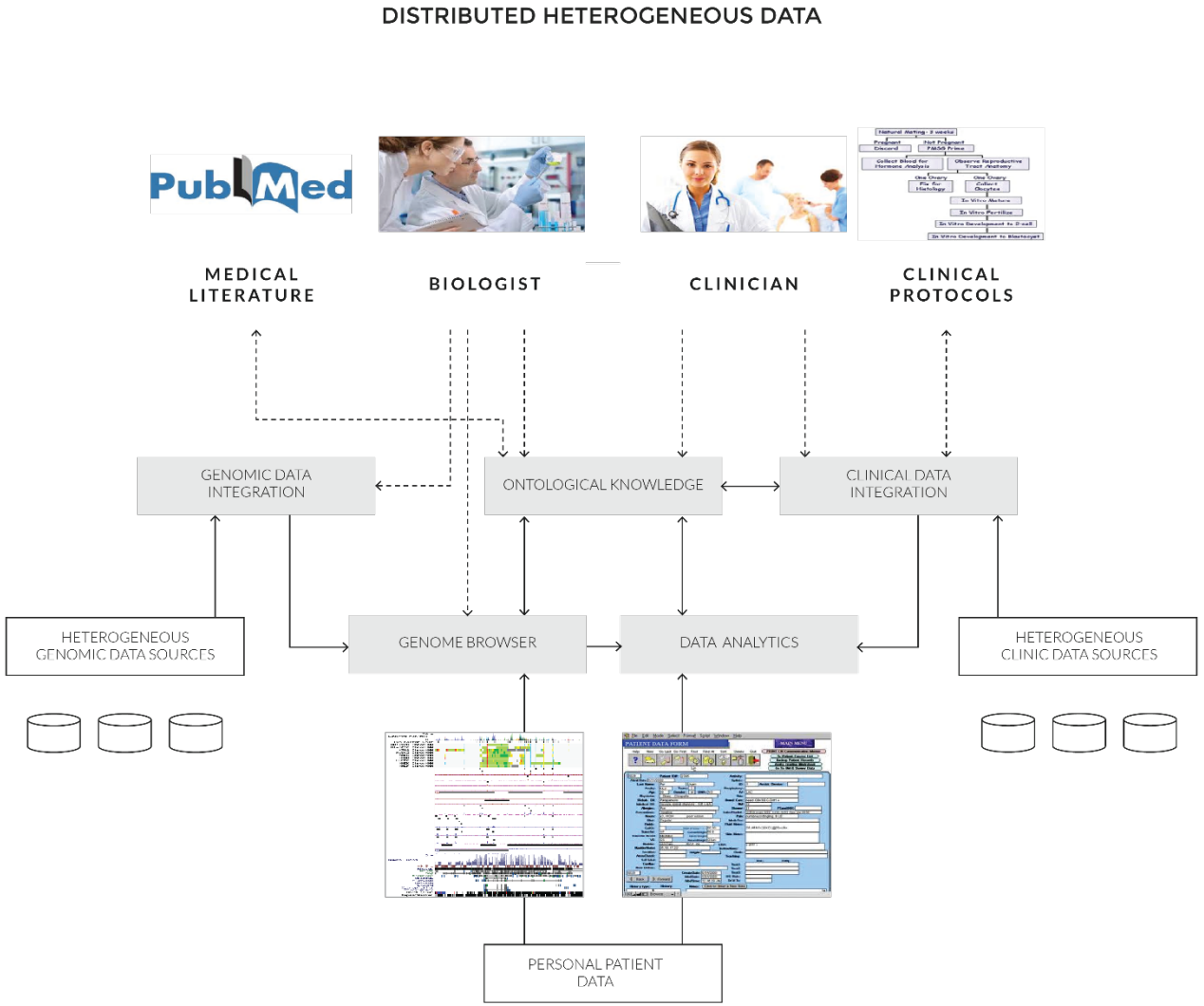
VISEO

Integrating structured and unstructured data

An integrated tool to monitor and automatically deal with user's complaints and interactions in social networks pages, and link this information with existing CRMs. Three advantages are envisaged:

- Centralize all users' information
- provide automatic, fast and personalized answers
- create valuable information about how customers respond to companies' products, helping companies in understanding their customers and improve their satisfaction.

Integrating text, numbers, logs and the like



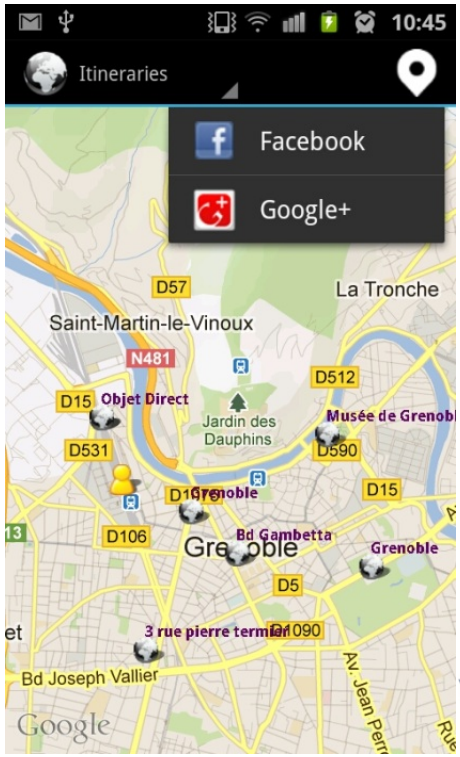
Integrating text, numbers, logs and the like

LEILAS
Language based Entity Identification for Location Aware Services

eurostars

Use of Natural Language Processing and semantic technologies to :

- Identify geographical named entities in text
- Link them with their spatial coordinates in order to increase precision in the identification and localisation phases



VISEO

Integration of approaches

Data science : the ultimate solution?

Heterogenous data : common methods?

- Symbolic methods?
- Machine learning?
- Hybrid methods?

VISEO

Scalability of the integrating ?

Will Salvation come from the Cloud ?

VISEO

Knowledge integration

Even with text only : NLP big challenge

While NLP technology is rather well advanced for sentence analysis, it is still in its infancy to relate information included in differents sentences, even more, in different documents

Knowledge integration

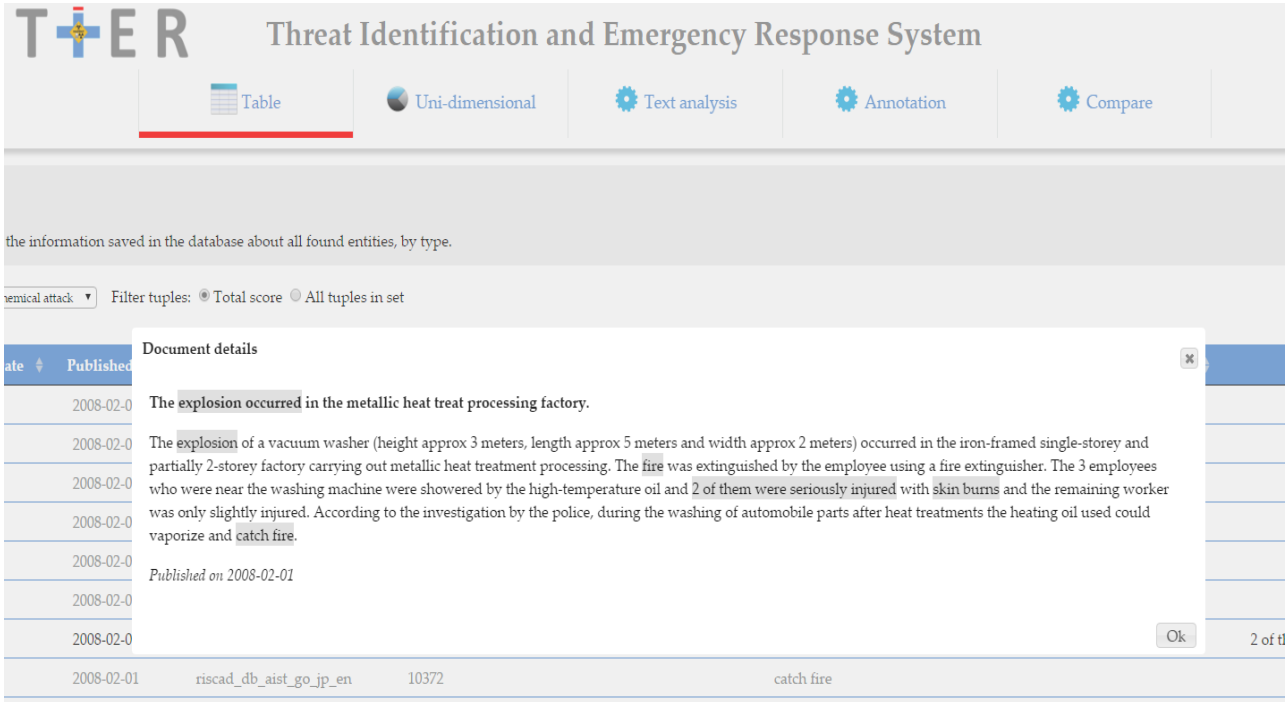
- When Apple unveiled the iPhone in January, Jobs hinted that Apple would be the only game in town for iPhone application development. He seemed concerned that a rash of third-party applications could create security and reliability problems that could derail Apple's first attempt at cracking the smart-phone market. (CNET News.com, June 12, 2007, 4:00 AM PDT)

  - ..to expand the capabilities of the iPhone so developers can write great apps for it but keep the iPhone secure," said Jobs, alluding to comments he made last week when he said security concerns precluded opening the phone to third-party software. (By Computerworld Published: June 11, 2007)

VISEO

Events extraction

The project aims at assessing the main risks and threats related to the potential malicious use of CBRN materials



VISEO

The « Big » challenge

Common representation of knowledge

Prepare the ingredients in such a way that they go well together :

- Need for a unified representation
- Common vocabularies, Ontologies

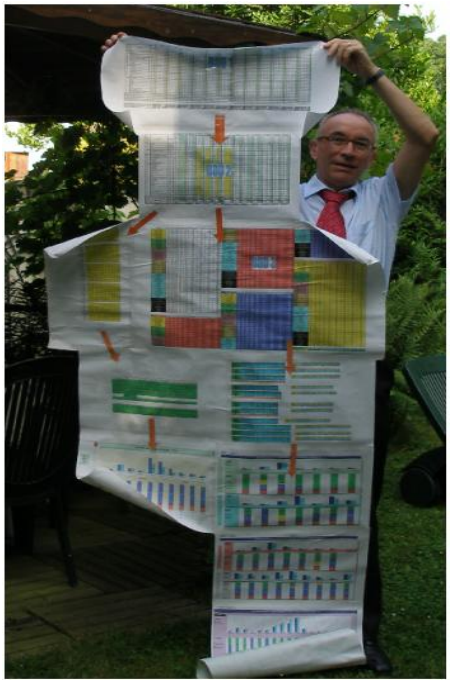
VISEO

Representing knowldege

How do we (still) do ?

Stones (Mesopotamia)

Graphs (Granada, 16 century)



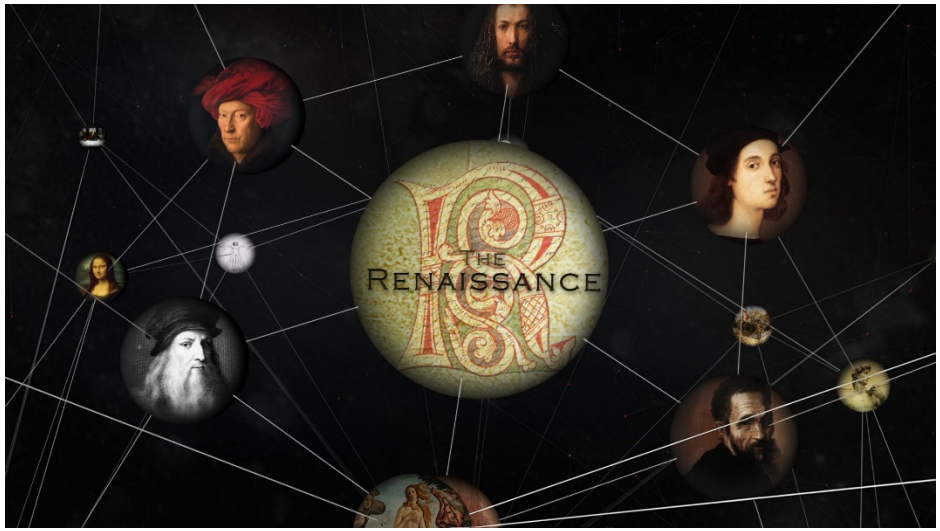
Multiple excel files

Stones, wood and sand

Courtesy of Jean Rohmer (Blog : PLEXUS LOGOS CALX)

VISEO

The big promises



Google knowledge graph

VISEO

Natural language ?

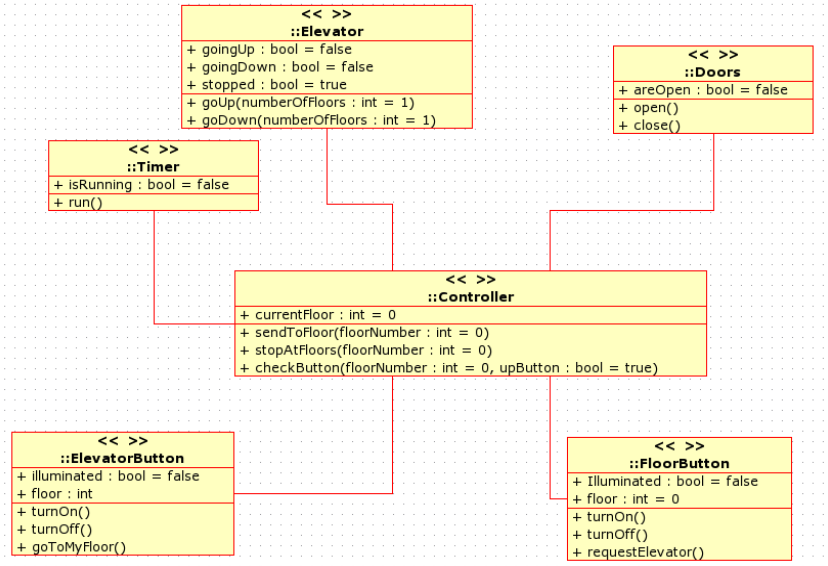
Is Natural language a solution ?

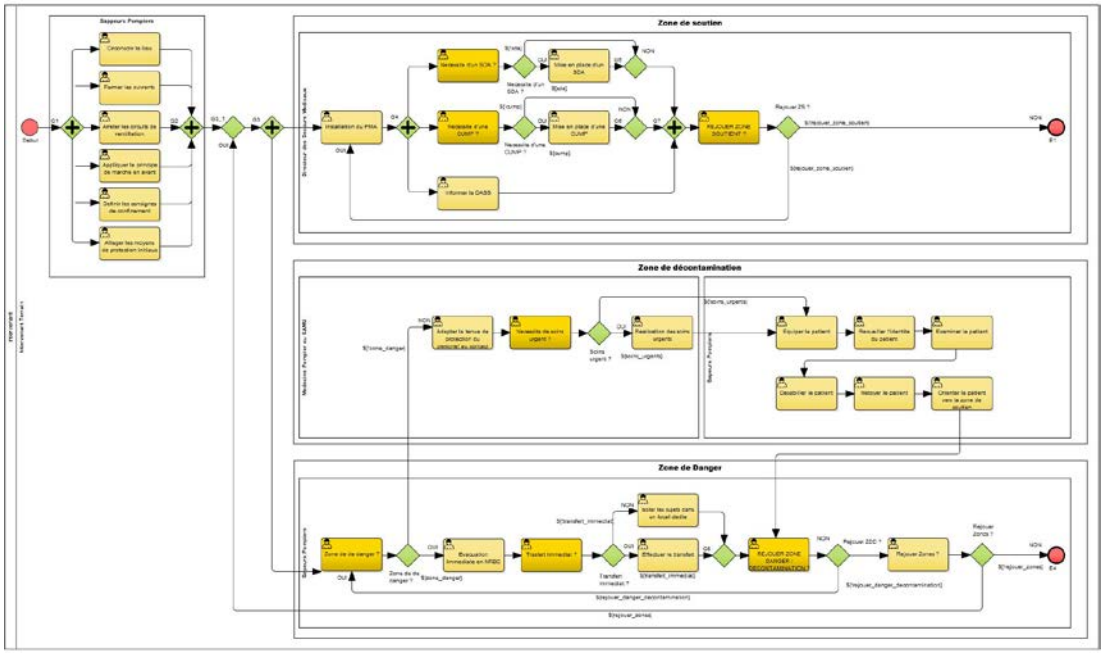
- Natural Language is the most natural way used to express knowldege, open source, exists since a long time and will exist a long time from now.
- But natural language is highly ambiguous

VISEO

Modeling Languages ?

UML, BPML?



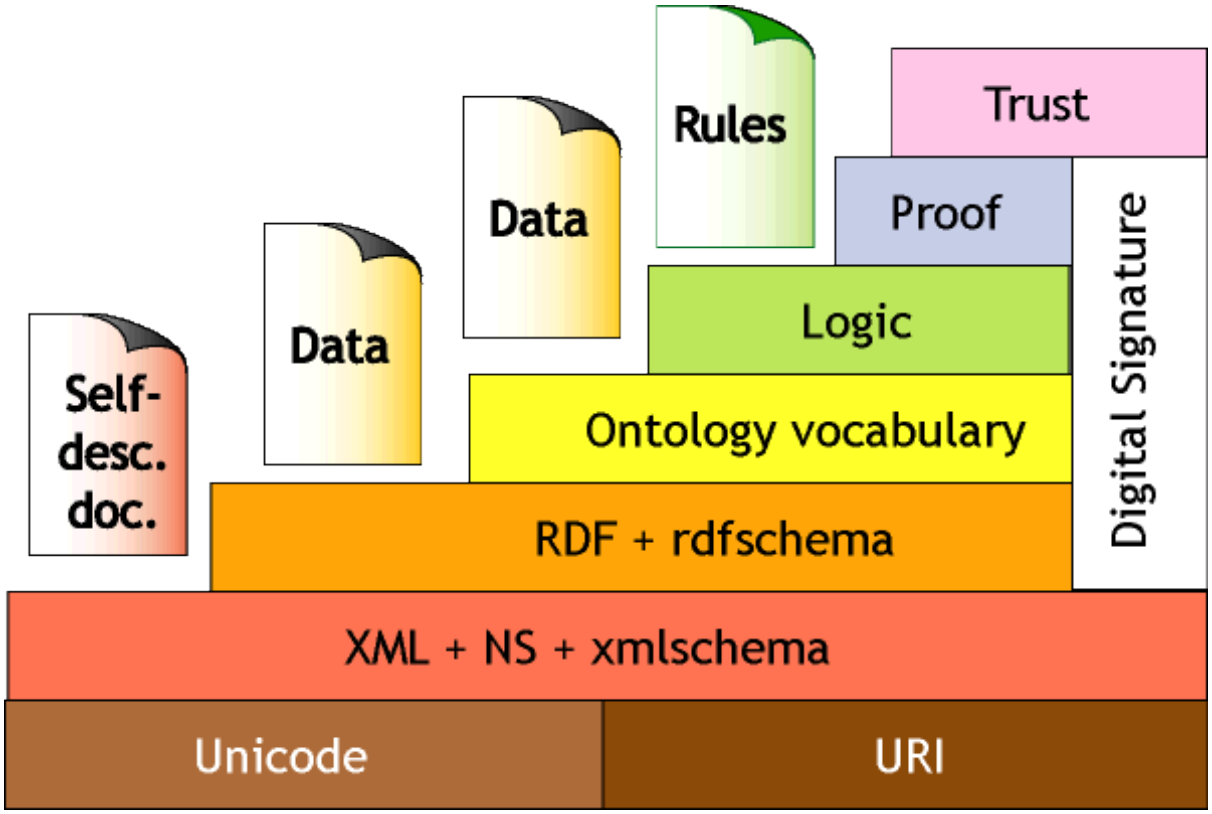


But no common vocabularies

VISEO

Semantic web ?

Is it the best pudding ?



Semantic Web cake layers

VISEO

SMILK

Joint Research and Innovation Laboratory

Combining NLP technologies with semantic web technologies to link knowledge contained on the Web

SMILK

Social Media Intelligence and Linked Knowledge

Agence Nationale de la Recherche

ANR

VISEO

Inria

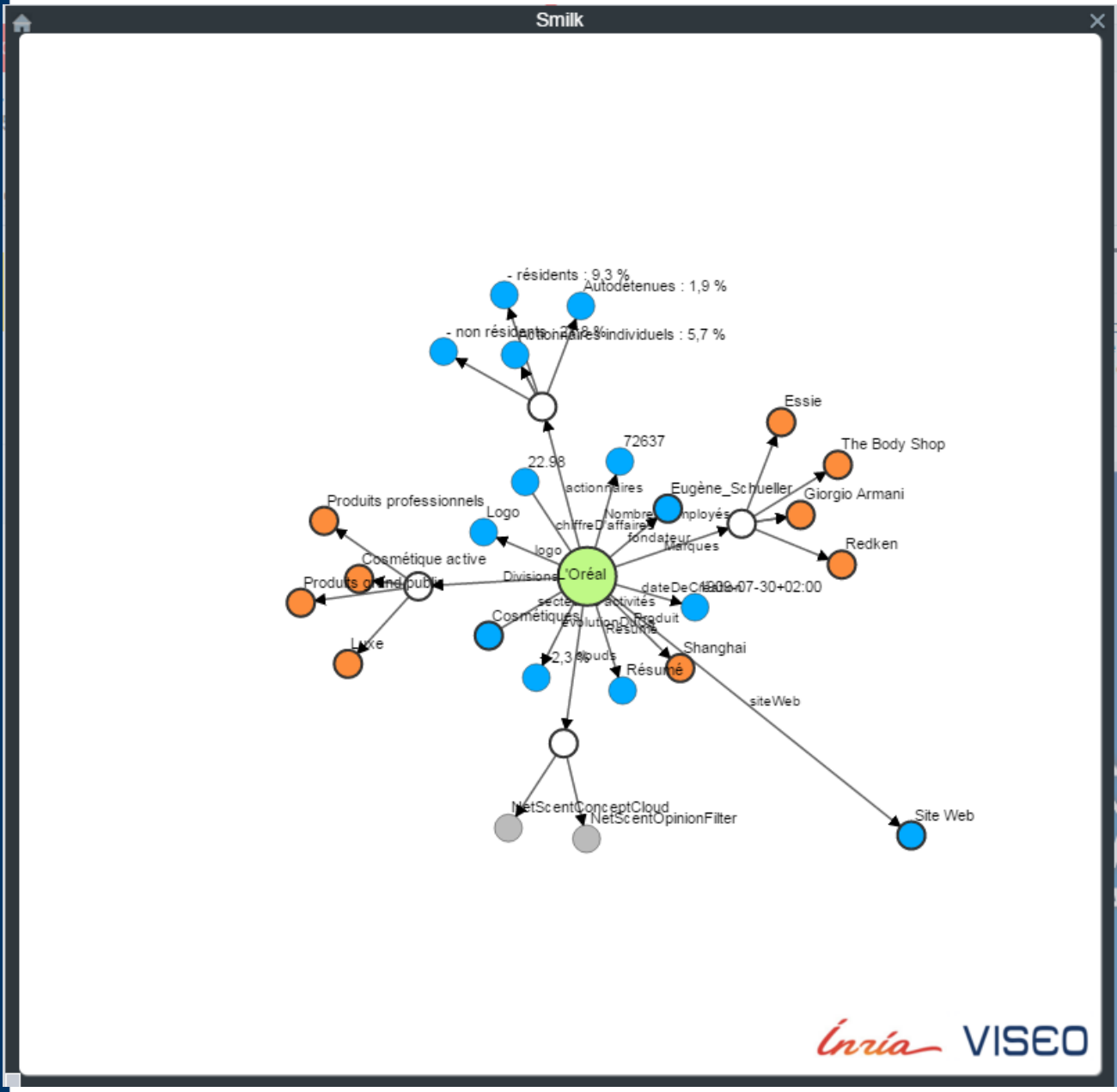
VISEO

SMILK

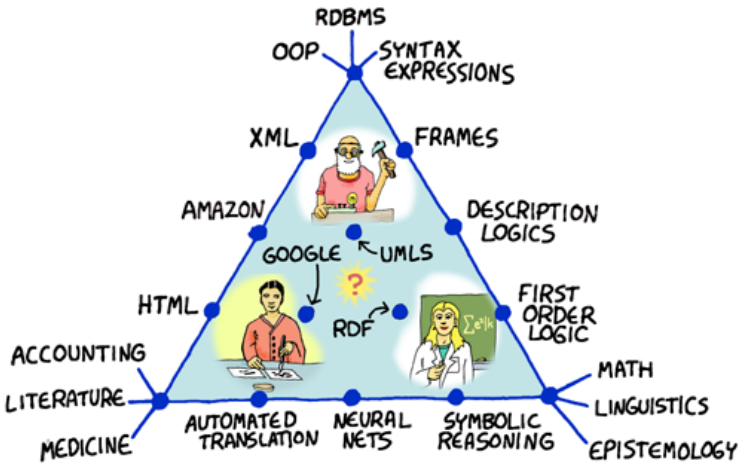

Économie française
confidentiels
sur www.sourcesure.eu
ÉCONOMIE
ÉCONOMIE FRANÇAISE
Loi Macron
Social
CAC 40
4 751.53 PTS
+1,2 %
DOW JONES(C)
17 168.61 PTS
-0,28 %
1 EURO
1.12 $US
-1,43 %
PÉTROLE
48.26 $US
+0,86 %
OR
1 163.95 $US
-0,17 %
La vie est belle pour Lancôme qui devance Dior et Chanel.
. LE MONDE ECONOMIE. | 10. 11. 2014 à 17h17. |. Par Juliette Garnier.
Abonnez vous à partir de 1 €
Réagir
Classer
. Partager (129). Tweeter.
LECTURE ZEN
L'Oréal a réussi son pari. A la barbe de LVMH, propriétaire de Dior, et de la marque Chanel, le groupe présidé par Jean-Paul Agon est parvenu à imposer un parfum Lancôme à la tête du palmarès des ventes de fragrances féminines en France. Depuis début 2014, La Vie est belle se range en pole position du marché hexagonal, d'après les chiffres de NPD arrêtés à fin septembre. Deux ans après son lancement, ce jus est devenu une référence. Chez Nocibé, chaque jour, il se vend 680 flacons, parfums et dérivés, de La Vie est belle. Le parfum dont la comédienne Julia Roberts est l'égérie a terrassé deux pointures : J'adore de Dior et N°5 de Chanel. " Du jamais vu depuis dix ans. Aucun nouveau parfum n'y était parvenu ", note Mathilde Lion, analyste du panel beauté chez NPD. " Une vraie réussite ", reconnaît Isabelle Parize, présidente de Nocibé. Dans son sillage, Si de.
L'accès à la totalité de l'article est protégé
Déjà abonné ? Identifiez-vous.
La vie est belle pour Lancôme qui devance Dior et Chanel.
Il vous reste 70% de l'article à lire
SMILK
Social Media Intelligence and Linked Knowledge
Certaines données n'ont pas pu être récupérées. Presser F5 pour réessayer.
Groupe
1. LVMH 1
Division
Marque
1. Dior 4
2. Chanel 4
3. Lancôme 3
4. Nocibé 2
Gamme
Produit
1. La Vie est belle 2
2. L'Oréal 1
3. N°5 1
Inria
VISEO

VISEO

SMILK


Smilk
- résidents : 9,3 %
Autodétenues : 1,9 %
- non résidents : 2,8 %
actionnaires individuels : 5,7 %
Essie
The Body Shop
72637
22.98
actionnaires
Eugène_Schueller
Giorgio Armani
Produits professionnels
Logo
chiffreDaffaires
Nombre
fondateur
marques
Redken
logo
Cosmétique active
Produits grand public
L'Oréal
dateDeCréation
2009-07-30+02:00
secteur
activités
Cosmétiques
Produit
Résumé
Luxe
2,3 %
Résumé
Shanghai
siteWeb
NetScentConceptCloud
NetScentOpinionFilter
Site Web

Conclusion



Unify our cooking efforts !

VISEO

Conclusion

To get the best out of the data pudding it is urgent to take care of

« *the Semantic warming* »

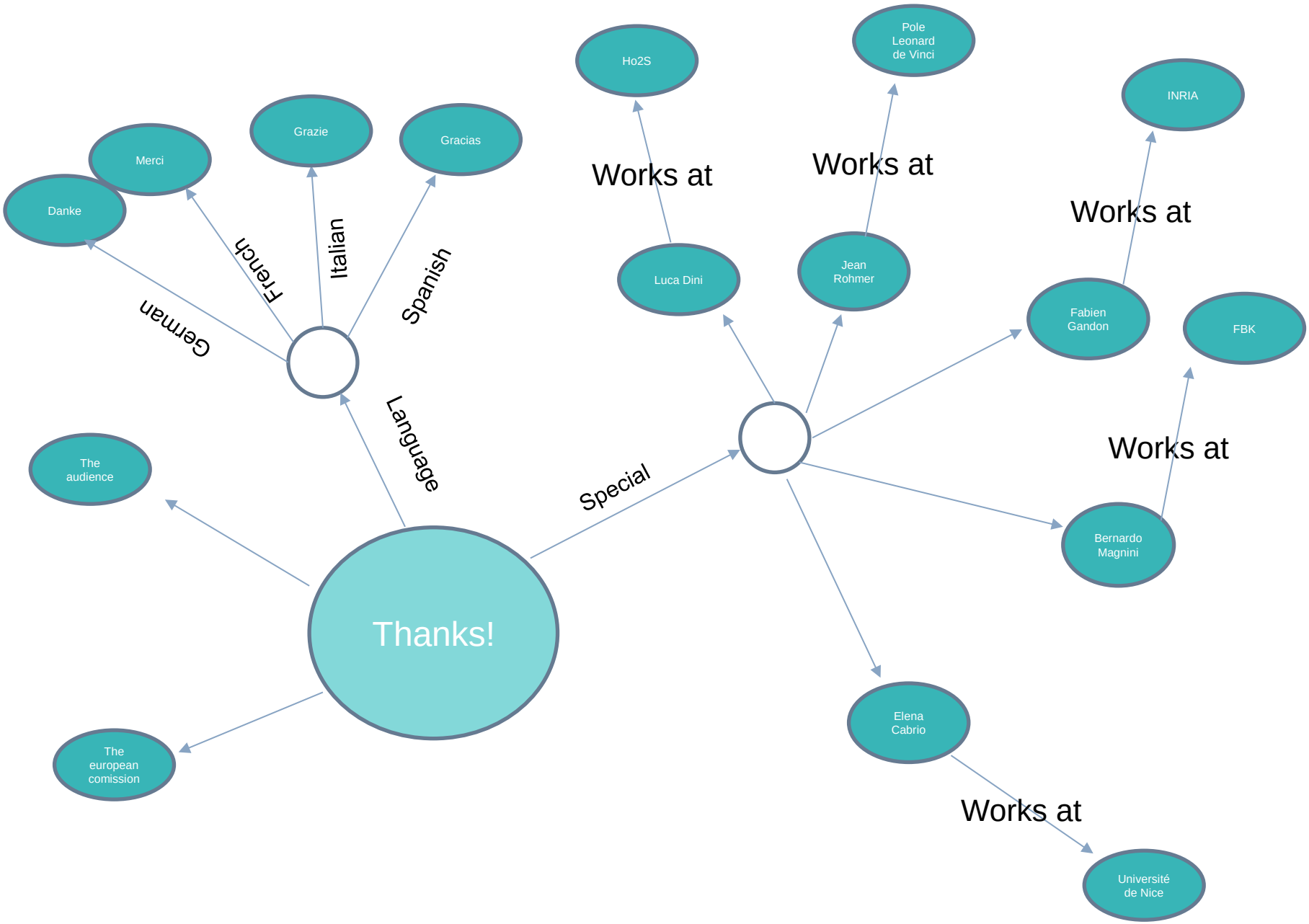
- Semantic processing
- Semantic electronic power
- Semantic promises

VISEO

Sources of images

- https://archive.4plebs.org/x/thread/15298566/
- http://www.dataminingblog.com/data-information-knowledge-and-wisdom/
- http://allrecipes.co.uk/recipe/25068/sticky-toffee-pudding-without-dates.aspx
- https://dataflog.com/read/how-big-will-the-internet-of-things-be/523
- http://lisperati.com/tellstuff/index.html
- http://enrouteglobalexchange.biz/social-network-sites-2.html
- http://www.actaddons.com/products/2005/call_on_crm.asp
- **http://www.bioinformatics.deib.polimi.it/genomic_computing/**

VISEO


Thanks!
Language
German
French
Italian
Spanish
Danke
Merci
Grazie
Gracias
Special
The audience
The european comission
Ho2S
Pole Leonard de Vinci
INRIA
Works at
Works at
Works at
Works at
Works at
Luca Dini
Jean Rohmer
Fabien Gandon
FBK
Bernardo Magnini
Elena Cabrio
Université de Nice