Analyse et modélisation des textes

Lou Burnard Consulting *octobre 2012*

1 Magritte nous rappelle

Qu'à chaque fois que l'on créée une représentation d'un texte ou d'autre chose, une image n'est pas réductible à la réalité.

2 Numériser et encoder…

La numérisation nous propose une *image*, une représentation d'un objet déjà existant
L'encodage nous permet de représenter l'image des *idées* résultant de ces représentations

Distinction fine mais très importante.

3 Des idées sur quoi?

Tout! mais on peut distinguer plusieurs axes…

- infos 'intrinsèques' à l'objet : les formes, couleurs, etc. d'une image; les sons, rythmes, etc. d'une musique; les structures linguistiques (mot, phrase, paragraphe) ou formels (chapitre, titre, nom de lieu) d'un texte…
- infos 'extrinsèques' ou 'meta' sur l'objet : son type, ses origines, ses buts, ses usages …
- infos 'interprétatives': la portée d'un texte ou d'un dessin, le programme d'une musique ou d'un rite…

distinctions floues mais pervasives…

L'essentiel est de faire le choix. Peut distinguer des actes, des informations intrinsèques à un objet, etc. Des informations dite méta, extrinsèques qui ne sont pas nécessairement présentes dans l'objet : ses origines, ses buts, etc. Enfin, des informations interprétatives.

4 Modélisation et structuration

Il existe (quelques) méthodes classiques d'analyses de données. L'important c'est de bien comprendre :

- toute méthode ne serait qu'une modélisation
- la modélisation devrait faire ressortir la structuration essentielle d'un objet complexe

Après une modélisation, on peut donner une implémentation informatisée ; sans modèle, une implémentation risque d'être incompréhensible et aléatoire, inutile…

Ce que l'on va vouloir faire, c'est modéliser. Pourquoi cette étape ? c'est qu'elle va nous permettre de donner une implémentation informatisée. Si vous n'êtes pas conscient de votre modèle, vous risquer d'avoir une implémentation inconsistante ou incompréhensible qui ne sera pas exploitable.

5 Analyse des données classique

On identifie…

- les 'objets d'interêt'
- leurs attributs/propriétés
- les relations entre objets
- les procédures/traitements essentiels envisagés

Dans l'analyse de données classique, on avait l'idée qu'il serait possible de conceptualiser l'ensemble de l'univers d'une entreprise. On identifie toute de suite les objets d'intérêt, leurs attributs et propriétés, etc.

On peut également appliquer ce modèle dans notre cas. Vous savez tous ce qu'est qu'une carte postale.

6 Analyse des documents

- quelles sont les unités que nous souhaitons traiter ?
- comment sont-elles structurées (quels composants, quels attributs) ?
- est-ce que les occurrences de ces objets sont clairement identifiables dans un flux textuel ?

Essayons cela avec ces trois documents :

Je vous propose avec l'exercice qui suit de identifier vous même un certain nombre de choses.

7 Document A

8 Document B

9 Document C

10 Exercice 2 : analyse d'un document

- Dans votre groupe, sélectionner d'abord un rapporteur
- Regarder bien votre document. Supposer que vous en avez quelques centaines d'autres pareils.
- Comment est votre document organisé ?
- Quels sont ses composants essentiels, commun à tous ces documents ?
- Quels sont ses composants intéressants, possible dans tous ces documents ?
- Sauriez vous faire en sorte qu'une autre personne reconnaîtrait les mêmes objets ?
- Faire une liste de tous les objets et propriétés essentiels de votre document
- Présenter cette liste… et justifier la !

L'une des leçons de cet exercice, c'est que l'on ne peut pas encoder mécaniquement un texte. Il est toujours intéressant de voir comment une personne envisage un document et se l'approprie.

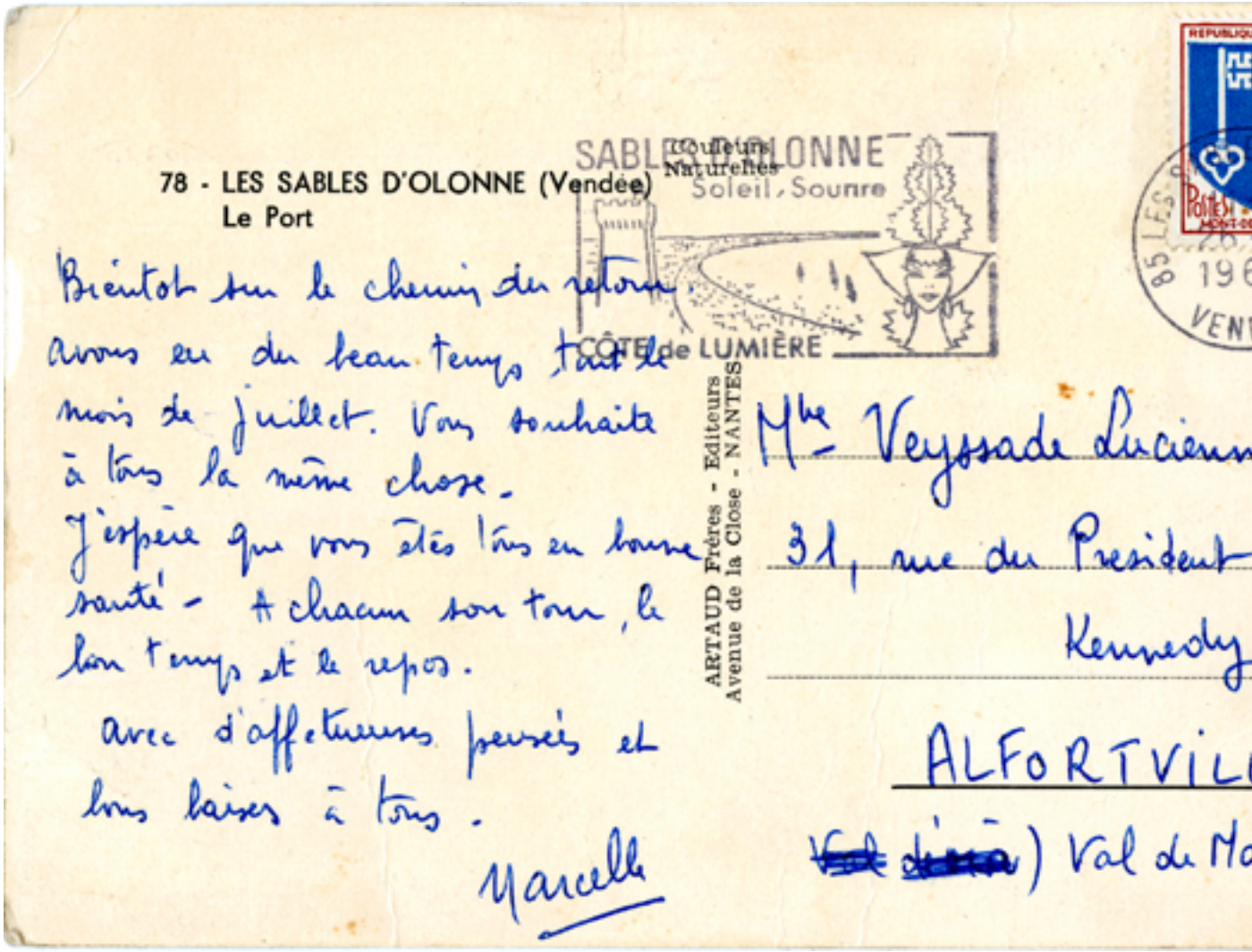

Figure 1: une carte postale

Mai 1763

ordre du Roy

le pavillon du Roy comme a l'ordinaire, mais l'on ne donnera de bois d'arrivée a personne

Le voyage etant comme celui de Versailles, l'on n'est obligé a rien, mais on peut leur donner quelque secours en bois et lumieres, mrs Bouchemen et le Moine oui suivement

oui sans doute

et ...
voy...
ordi...
Si...
Le...
Gar...
des...
et ...
Si...
de...
la ...

F. LE LIONNAIS
23, route de la Reine

92 - BOULOGNE SUR SEINE

Tél. 605 90 13

le 3 Mars 1969

Chers Brigadiers,

Le sort en reste jeté. Raymond QUENEAU et René THOM viennent d
firmer - avec des accents qui ne laissaient aucun doute sur le
rité - que notre prochaine réunion aura bien lieu :

Lundi 17 Mars à 17 Heures, chez Georges PEREC (4° Etage)
92, rue du Bac PARIS 7° - Téléphone : 222 95 68

On est prié de ne pas arriver en retard, car certains particip
exprimé leur désir de se libérer avant 18 H 45.

Il avait récemment couru le bruit que le signataire de cette l
coutumier du parjure et qu'on ne pouvait lui faire confiance p
vraison du "Deuxième Manifeste" destiné aux Editions Cape. C'é
pure calomnie. Voici le texte ou du moins un projet que vous ê
à analyser, critiquer et compléter. Voyez-vous d'autres aspect
OuLiPien qui mériteraient d'être mentionnés ? Ou croyez-vous q
ration sémantique devrait être développée ?

Je vous laisse sur ces méditations et vous adresse mes pensées

P.S. Claude BERGE : Et bien entendu, je t'attends chez moi Jeudi 13
11 Heures.

Georges PEREC : THOM aimerait beaucoup avoir un tableau noir.
vous vous le procurer ?

Jacques BENS : Merci pour votre lettre du 25 Février. L'idée
distinction entre "Lipo pure" et "Lipo appliquée" me semble d'
certaine. Il vaudrait certainement la peine de s'y intéresser
cela donnerait-il un alinéa de plus dans le Manifeste ci-joint

Venez-vite (et restez longtemps), on sera content de vous revo