# Titre : Organisation, agrégation et visualisation d'informations médiatiques

Colloque « Fonder les sciences du territoire », 23-25 novembre 2011, Paris

**AUTEURS**

Robin Lamarche-Perrin, Yves Demazeau et Jean-Marc Vincent

**RÉSUMÉ**

Cet article présente une méthode de traitement de données pour GEOMEDIA, une plateforme de visualisation et d'analyse d'informations médiatiques. Devant la complexité et la quantité des données médiatiques, nous avons trois objectifs. (1) Organisation du déluge d'informations pour en extraire la structure générale : trois dimensions (Agents × Dates × Thèmes) sont retenues pour classer les articles. Nous utilisons la notion d'« agent », issue de l'Intelligence Artificielle, pour généraliser la notion d'« espace ». (2) Agrégation de l'information pour obtenir un point de vue macroscopique sur la structure de données : des procédures automatiques d'agrégation réduisent la complexité structurelle et engendrent ainsi des abstractions de haut-niveaux. (3) Visualisation des données agrégées : projections spatiales sur des cartes géométriques, territoriales ou réticulaires, projections temporelles sur des frises chronologiques, projections thématiques sur des nuages sémantiques, *etc.*

**ABSTRACT**

This article presents some data processing technics for GEOMEDIA, a platform for visualization and analysis of media information. In order to handle the complexity and the large amount of media data, we proceed in three stages. (1) Organization of the information flow in order to extract its global structure: Three dimensions (Agents x Dates x Themes) are settled for data classifying. We use the concept of "agent", originated from Artificial Intelligence, in order to generalize the concept of "space". (2) Aggregation of information in order to obtain a macroscopic point of view: Aggregation processes automatically decrease the structural complexity of data. Thus, they generate high-level abstractions. (3) Visualization of aggregated data: spatial projections on geometric, territorial or reticular spaces, temporal projections on timelines, thematic projections on tag clouds, *etc.*

**MOTS CLÉS**

Traitement de données médiatiques, espace et temps, analyse macroscopique

## INTRODUCTION

Les médias internationaux engendrent chaque jour un très grand nombre de données hétérogènes. Cet article fait partie du programme GEOMEDIA visant à la conception d'une plateforme de visualisation des informations médiatiques. Le projet est coordonné par Claude Grasland, Timothée Giraud et Marta Severo (*cf.* leur communication lors de ce colloque). Nous présentons ici des méthodes issues de l'Intelligence Artificielle, plus particulièrement du domaine des « systèmes multi-agents » (Wooldridge, 2002), pour traiter le déluge d'informations en provenance des médias. L'objectif est d'analyser le contenu de l'actualité et le rapport entre ce contenu et les médias concernés.

Trois difficultés sont abordées :

1. Comment organiser les données malgré leur hétérogénéité ? La section 1 pose trois dimensions à partir desquelles le déluge d'informations peut être structuré : un espace d'agents (qui généralise la notion d'espace classique), une dimension temporelle et une dimension thématique.
2. Comment appréhender de très grandes quantités de données ? La section 2 présente des opérations d'agrégation destinées à réduire la complexité structurelle du déluge. Les abstractions ainsi engendrées offrent une sémantique macroscopique pour la description et l'analyse des informations.

3. Comment visualiser les données médiatiques ? La section 3 explique comment les trois dimensions peuvent être projetées sur des interfaces de visualisations : projections spatiales sur des cartes géométriques, territoriales ou réticulaires, projections temporelles sur des frises chronologiques, projections thématiques sur des nuages sémantiques, *etc.*

## 1. ORGANISATION DES DONNÉES

### Les trois dimensions de l'information

Trois dimensions sont retenues pour classer les articles de presse : espace, temps et thème. L'espace et le temps sont retenus pour l'importance primordiale des lieux et des dates dans l'analyse de l'actualité. Ces deux dimensions structurantes (la base du modèle) sont complétées par un axe sémantique : la dimension thématique. Elle est modélisée avec très peu de contraintes, afin de garantir la généricité du modèle final. Enfin, l'espace et le temps peuvent organiser le *contenu* d'un article (information véhiculée) ou le *contenant* (source de l'information) :

1. **L'espace** d'un article regroupe les lieux relatés par l'information (contenu) ou les lieux d'où provient l'information (contenant).
2. **Le temps** d'un article regroupe les dates et intervalles de temps relatés par l'information (contenu) ou les dates de rédaction et de publication de l'article (contenant).
3. **Le thème** d'un article regroupe les catégories thématiques de l'information véhiculée (politique, économique, sportif, *etc.*). Même si elle peut être combinée avec les caractéristiques spatio-temporelles du contenant (*e.g.*, variation des thèmes en fonction du lieu et de la date de publication), la dimension thématique caractérise toujours le *contenu* de l'information. Il s'agit d'une *dimension sémantique* (et non *structurelle*).

Aucun article ne semble manquer simultanément à ces trois catégorisations. Nous négligeons les cas marginaux d'informations médiatiques non-localisées et/ou atemporelles. De plus, au niveau du contenant, un article est *toujours* caractérisé par un lieu et une date de publication. D'autres dimensions pourraient être retenues pour organiser les informations médiatiques, notamment pour organiser la *signification* de l'information (*i.e.*, « *ce qui y est dit* » : les faits, les opinions et arguments déployés, leur analyse, *etc.*). Ce projet s'intéresse cependant à la *structure* globale du déluge médiatique, plus qu'aux sémantiques des informations. L'espace et le temps constituent ainsi deux dimensions structurantes fondamentales. En outre, la dimension thématique est suffisamment générale pour encadrer des regroupements lexicaux particuliers et des analyses sémantiques simples.

### De l'espace géographique à l'espace des agents

Une originalité de ce projet réside dans la modélisation de l'espace en termes d'agents. Un *agent* est une abstraction pouvant modéliser toute entité proactive et autonome (Wooldridge, 2002) : un individu, une administration, une collectivité, un État, un gouvernement, *etc.* Nous substituons donc à l'espace géographique classique un « espace d'agents ». Il généralise l'espace territorial (un agent peut être un territoire) et enrichit la dimension spatiale. L'objectif de cette généralisation est d'offrir un modèle générique pour d'autres espaces : administratifs, politiques (partis, syndicats), virtuels (sites internet, blogosphère), *etc.* La section 2 montre comment cette généricité permet de redéfinir les espaces médiatiques en apportant de nouveaux maillages pertinents pour l'analyse géographique. La section 3 montre comment ces « espaces d'agents » peuvent être visualisés en les projetant sur des espaces géographiques ou territoriaux classiques.

### Collecte et organisation des données

La collecte des données utilise la technique d'agrégation de flux RSS présentée par Claude Grasland, Timothée Giraud et Marta Severo (ce colloque). Trois dictionnaires sont utilisés pour indexer le contenu des articles recueillis en fonction des trois dimensions : (1) un dictionnaire d'agents, (2) un dictionnaire de dates et (3) un dictionnaire de thèmes. Chacun d'eux associe donc un agent, une date ou un thème à un ensemble de mots-clés. La coprésence de plusieurs mots-clés dans un même article est ensuite utilisée pour placer celui-ci dans la « structure multi-agents » (structure à trois dimensions : Agent × Date × Thème).

- Un triplet (Agent, Date, Thème) définit un *évènement* répertorié à une date donnée dans l'histoire de l'agent. *E.g.*, (« Fukushima », « mars 2011 », « énergie atomique ») témoigne d'un évènement relatif au nucléaire dans la préfecture de Fukushima en mars 2011.
- Un quadruplet (Agent1, Agent2, Date, Thème) définit une *interaction spatiale*, liant deux agents à un instant donné. *E.g.*, (« FC Barcelone », « Manchester United », « 28 mai 2011 », « Ligue des champions »).
- Un quadruplet (Agent, Date1, Date2, Thème) définit une *interaction temporelle*, liant deux dates pour un agent donné. *E.g.*, (« Élysée », « 1981 », « 2012 », « présidentielles »)
- Un quintuplet (Agent1, Date1, Agent2, Date2, Thème) définit une *interaction spatio-temporelle*, liant deux agents à des dates distinctes. *E.g.*, (« USA », « 11 septembre 2001 », « Pakistan », « 2 mai 2011 », « terrorisme ») témoigne d'un lien entre des évènements relatifs à la lutte anti-terroriste, ayant eu lieu en 2001 aux USA et en 2011 au Pakistan.

- Des relations de coprésence plus simples (couples d'agents, de dates ou de thèmes), ou plus complexes, peuvent être positionnées de la même manière dans la structure multi-agents. Elles sont toujours interprétées comme des relations entre des agents et/ou des dates autour de thématiques particulières.

Il est possible d'utiliser le contenant des articles pour les positionner. Les agents correspondent alors aux médias eux-mêmes et les dates à la rédaction ou publication des articles. La structure obtenue n'a pas la même signification. Au lieu de représenter des relations entre agents de manière neutre par rapport aux médias analysés, elle représente les relations « entre un média et un agent » ou « entre une date de publication et un évènement relaté ». Les deux approches sont potentiellement utiles, mais elles n'amènent pas à travailler sur les mêmes objets : « le contenu de l'actualité », d'une part, et « le rapport entre ce contenu et les médias responsables », d'autre part.

**Diagrammes multi-agents**

Les structures multi-agents peuvent être représentées sous forme de diagrammes : le temps en abscisse (ensemble de dates), l'espace des agents en ordonnée et la dimension thématique est représentée par un ensemble de couleurs. Un *évènement* est représenté par un point coloré de l'espace-temps (en gris dans la Fig. 1). Une interaction est représentée par un segment coloré (horizontal pour une *interaction temporelle* (Fig. 2), vertical pour une *interaction spatiale* (Fig. 3), diagonal pour une *interaction spatio-temporelle* (Fig. 4)). Enfin, l'épaisseur des points et des segments est relative à la quantité d'articles qui y font référence. Voici les diagrammes correspondant aux exemples simples présentés ci-dessus : (Il s'agit d'illustrations simples présentant très peu d'agents, de dates et de relations. Bien évidemment, plusieurs types de relations peuvent apparaître *dans un même diagramme*).

Fig 1. Évènements

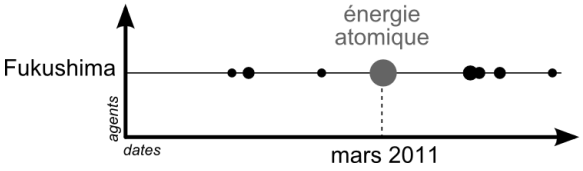



Fig 2. Interactions temporelles

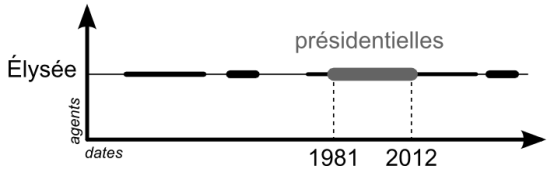



Fig 3. Interactions spatiales

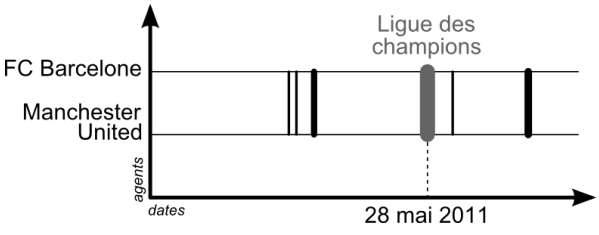
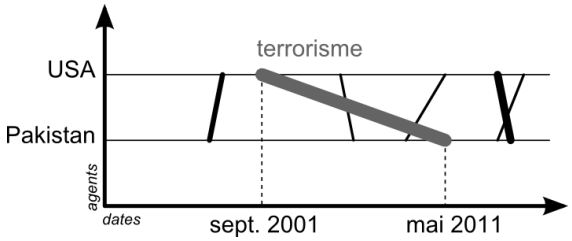



Fig 4. Interactions spatio-temporelles




## 2. AGRÉGATION DES DONNÉES

La plateforme GEOMEDIA est destinée à traiter de très grandes quantités de données. L'objectif des fonctions d'agrégation présentées dans cette section est de généraliser les informations médiatiques pour produire des abstractions macroscopiques utiles.

**Définition**

Dans nos travaux précédents, une structure multi-agents est *macroscopique* par rapport à une autre lorsque *sa complexité est inférieure* (Lamarche-Perrin, Demazeau, Vincent, 2011). La notion de complexité développée est volontairement générique et relative. Elle mesure la difficulté à manipuler les données d'une structure. Elle est donc liée à la quantité d'informations manipulées selon une procédure d'investigation donnée (Klir, 1985). Le nombre d'agents, le pas temporel, le nombre d'évènements ou d'interactions, sont autant de mesures de complexité simples. Les fonctions d'agrégation minimisent ces mesures en modifiant les structures multi-agents. Elles entraînent de fait une perte d'information, mais permettent d'abstraire les données et d'élaborer une description macroscopique. L'objectif est alors de maximiser la quantité d'informations contenues dans la structure macroscopique, tout en minimisant sa complexité.

**Fonctions d'agrégation**
Un *agrégat* est un sous-ensemble de la structure multi-agents : agrégat spatial (ensemble d'agents), agrégat temporel (ensemble de dates, période de temps), agrégat thématique (lexique, ensemble de thèmes) ou agrégat mixte (ensemble d'évènements, d'interactions, *etc.*). Une *agrégation* remplace la structure d'un agrégat par une structure moins complexe. Afin de garantir la cohérence de l'image macroscopique vis-à-vis des données exploitées, les agrégations doivent respecter certaines contraintes. Celles-ci sont généralisées à partir des travaux de Mattern (1989) sur les « systèmes distribués » (*cf.* (Lamarche-Perrin, Demazeau, Vincent, 2011) pour plus de détails). Les trois contraintes qui suivent sont destinées à garantir la correspondance entre les abstractions de haut-niveau et les données initiales. Elles sont essentielles, notamment afin de conserver les relations de causalité spatiales et temporelles au sein de la structure multi-agents.

1. **Fermeture.** La structure substituée ne doit pas présenter d'éléments nouveaux (évènements ou interactions) : par exemple, pour un ensemble d'agents remplacé par un agent virtuel unique, les évènements et interactions de l'agent virtuel doivent correspondre à des évènements et interactions des agents de l'agrégat.
2. **Complétude.** La structure substituée ne doit pas omettre d'éléments de l'agrégat : les évènements et interactions des agents de l'agrégat doivent être représentés par des évènements et interactions de l'agent virtuel. Par contre, ils peuvent éventuellement être agrégés entre eux (*e.g.*, plusieurs évènements sont représentés par un seul et unique évènement).
3. **Cohérence.** L'ordre des éléments de la structure substituée doit correspondre à l'ordre des éléments de l'agrégat. Notamment, les relations de causalité entre les évènements ou entre les interactions doivent être conservées lors de l'agrégation.

Le respect de ces contraintes garantit que la structure macroscopique correspond aux informations microscopiques à partir desquelles elle est engendrée. De plus, pour maintenir cette correspondance, des agrégations sont souvent répercutées, notamment sur les relations complexes de la structure multi-agents. Ainsi, un *évènement* (Agent, Date, Thème) est agrégé avec un autre évènement si ses trois composantes sont agrégées unes-à-unes. Par exemple, des « manifestations » à Benghazi le 15 février 2011 et des « affrontements » à Tripoli le 25 février 2011 peuvent être perçus comme un seul est même évènement pour une agrégation aux échelles nationale et mensuelle, sous les catégories thématiques générales de « contestations » ou de « révolutions » (*cf.* diagrammes ci-dessous).

Fig 5. Structure initiale

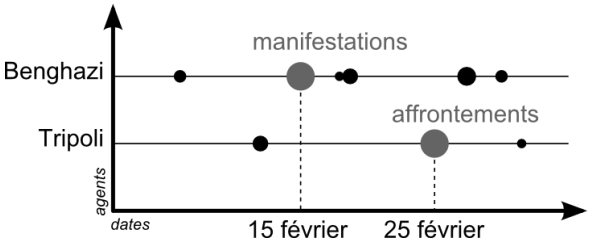
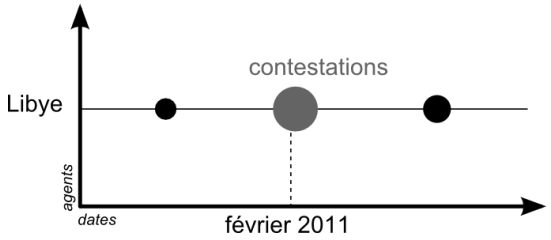



Fig 6. Structure macroscopique




**Interprétation des agrégats**
Parmi la multitude des agrégations possibles, les agrégats retenus sont ceux qui réduisent efficacement une mesure de complexité donnée. Ils généralisent ainsi l'information microscopique et engendrent des abstractions de haut-niveau. Par exemple, en agrégeant des agents qui partagent beaucoup d'interactions (Agent1, Agent2, Date, Thème) ou des périodes de temps durant lesquelles les interactions sont également très fréquentes (Agent, Date1, Date2, Thème), on réduit le nombre de ces interactions et la mesure de complexité associée. Les agrégats sont interprétés comme des « proximités médiatiques » entre des agents, des périodes et des thématiques. Ils sont élaborés automatiquement à partir de la structure même de l'information. Ils y définissent des catégories nouvelles, éventuellement difficiles à nommer, mais utiles à l'analyse dans la mesure où elles offrent un regard macroscopique (*i.e.* elles réduisent la complexité) en épousant la structure même du système (*i.e.* les interactions entre ses éléments).

## 3. VISUALISATION DES DONNÉES

Chaque agent est caractérisé par un ensemble de lieux, soit de manière triviale (l'agent « France » est caractérisé par la localisation territoriale « France »), soit par convention (les agents « Assemblé nationale française » et « président de la République française » sont également caractérisés par la localisation « France »). L'« espace des agents » peut ainsi être projeté sur un « espace réel » de plusieurs façons différentes : à l'aide de coordonnées GPS par

exemple, pour une projection sur un « espace géométrique » ; en fonction de l'appartenance à une région ou à un État, pour une projection sur un « espace territorial » ; en fonction de la place de la dépêche au sein d'un réseau d'information, pour une projection sur un « espace réticulaire ». Ces projections permettent de visualiser l'organisation spatiale de la structure multi-agents sur différents types de cartes ou de graphiques. De la même manière, l'espace des dates peut être visualisé sur des frises chronologiques : à temps continu (dates) ou discontinu (intervalles), à temps linéaire ou non-linéaire, avec des cycles (sur la semaine, le mois, l'année), *etc.* Des travaux en « logiques temporelles » peuvent encadrer ces projections (Allen, 1983). Enfin, les thèmes peuvent être projetés sur des espaces sémantiques (*e.g.*, nuages de mots-clés) grâce à des techniques d'analyse textuelle.

Les agrégations de la structure multi-agents induisent des rapprochements au niveau des interfaces de visualisation. Les agrégations d'agents, par exemple, amènent à rapprocher des lieux et des territoires, à constituer des groupements de pays en fonction des interactions entre leurs agents. Les agrégations de dates modifient le pas temporel de la représentation, la granularité de certaines périodes, le détail de certains évènements, *etc.* Les cartes et les frises chronologiques sont donc *directement* modifiées par les processus d'agrégation. Ces interfaces ne sont pas un cadre au sein duquel on représente la structure de l'information, elles sont elles-mêmes structurées *comme l'information*. Si bien qu'en un coup d'œil, on peut saisir les caractéristiques macroscopiques du déluge d'informations, selon ses trois dimensions structurantes (espace, temps et thématiques) ou selon le croisement de ces dimensions (évènements, interactions spatiales, temporelles, *etc.*).

## 4. PERSPECTIVES

**Implémentation du modèle**
La prochaine étape du projet est la réalisation d'un prototype de plateforme GEOMEDIA. Celui-ci implémentera les fonctions présentées dans cet article (organisation, agrégation et projection des données). Il servira à faire une pré-évaluation des méthodes de généralisation par agrégation de données. Des expériences préliminaires sur un ensemble de 40 000 articles sont en cours. Elles seront présentées lors du colloque.

**Mise en pratique du modèle**
L'utilité effective du modèle présenté sera ensuite évaluée par une mise en pratique de la méthode d'analyse macroscopique. Nous nous concentrerons sur des acteurs, une période et un thème qui restent à déterminer (*e.g.*, les politiques énergétiques des pays européens après la catastrophe nucléaire de Fukushima). Les résultats des agrégations seront interprétés et évalués en fonction de leurs pouvoirs descriptif et explicatif. L'approche sera considérée comme pertinente si les outils facilitent le travail d'analyse d'un initié ou d'un expert. Elle sera considérée comme innovante si elle engendre des abstractions nouvelles et utiles pour l'explication du sujet abordé.

## REFERENCES


Allen J. F., 1983, « Maintaining knowledge about temporal intervals », *Communications of the ACM*, 26 (11), nov. 1983, ACM, New York, USA, p. 832-843.

Klir G. J., 1985, « Complexity: Some general observations », *Systems Research*, 2 (2), juin 1985, John Wiley & Sons, Ltd., p. 131-140.

Lamarche-Perrin R., Demazeau Y., Vincent J.-M., 2011, « Observation macroscopique et émergence des systèmes multi-agents de très grande taille », *19e Journées Francophones des Systèmes Multi-Agents*, 17-19 oct. 2011, Cépaduès, Valenciennes, France, p. 53-62.

Mattern F., 1989, « Virtual Time and Global States of Distributed Systems », *Parallel and Distributed Algorithms*, Elsevier, North-Holland, p. 215-226.

Wooldridge M., 2002, *An Introduction to MultiAgent Systems*, John Wiley & Sons, Ltd., Chichester, Angleterre.


## LES AUTEURS


Robin **Lamarche-Perrin**
Université de Grenoble
robin.lamarche-perrin@imag.fr

Yves **Demazeau**
CNRS, Grenoble
yves.demazeau@imag.fr

Jean-Marc **Vincent**
Université Joseph Fourier, Grenoble
jean-marc.vincent@imag.fr