PLATE-FORME POUR L'INDEXATION SPATIALE MULTI-NIVEAUX D'UN CORPUS TERRITORIALISÉ

par Julien Lesbegueries

post-doctorant au LSIIT, ULP Strasbourg
E-mail : julienlesbegueries@gmail.com

Notre travail s'insère dans la problématique de l'accès à l'information spatiale présente dans des corpus textuels territoriaux. Nous proposons d'aller au-delà des systèmes de recherche d'information classiques, — basés sur l'analyse statistique des document et peu adaptés à ce cas particulier—, via un traitement linguistique ciblé interprétant l'information spatiale. Notre hypothèse est que des traitements relativement peu coûteux suffisent à dégager l'essentiel de l'information. Ils sont un bon point de départ pour une interprétation plus poussée en utilisant les propriétés géographiques de l'information extraite et le développement d'un système d'indexation à plusieurs niveaux d'abstraction. Nous proposons en effet une méthode de recherche d'information spatiale multi-niveaux indexant un corpus textuel brut. Cette méthode qui extrait l'information d'un corpus et l'interprète, permet d'améliorer l'efficacité des systèmes de recherche d'information à chaque fois que l'interrogation comporte une connotation spatiale. L'interprétation permet en outre de retrouver le contexte dans lequel l'information spatiale a été utilisée. En particulier, elle permet d'indexer des unités de texte en leur associant des contextes de type itinéraire, description locale ou comparaison de lieux.

The aim of our work is to provide an easier way to access documents in territorial corpora and, particularly, spatial information contents. We suggest to go further than classical based-statistic information retrieval systems that are not suitable in the case of spatial information extraction. A light linguistic process can rather be used in order to draw the information's main thing. They can be a good starting point to be used thereafter in a more precise interpretation process, using only the geographic properties extracted, in order to propose a multi-level indexing method, each level corresponding to an abstraction level of spatial information. Thus, we propose a multi-levels spatial information retrieval system, indexing unstructured textual documents. This method, that interprets spatial information, allows to improve the efficiency of information retrieval systems each time a spatial query is performed. This interpretation can also retrieve the context in which the spatial information is used by the author. Particularly, text units can be classified in ''itinerary", ''local description" or ''area comparison" contexts.

1 Introduction – Définition de l'information spatiale dans le texte

L'expression de l'information spatiale est une problématique abordée dans divers domaines et, particulièrement, par les linguistes (Vandeloise 1986 ; Denis 1997) et les cognitiens (Cohn 1997 ; Egenhoffer 1991). Ceux-ci ont essayé de formaliser la manière dont l'homme se représente l'espace qui l'entoure et ont proposé des modèles :

- Le modèle cible/site de Vandeloise explique le concept de cible comme étant l'entité spatiale décrite à l'aide d'une entité dite « site », supposée connue (La voiture (cible) est près de la maison (site)). Cette proposition fait l'hypothèse que l'information spatiale est donc toujours relative à un point de repère.

- Le modèle RCC9 de Cohn et la matrice 9-intersections d'Egenhoffer sont des modèles plus poussés du raisonnement spatial qualitatif, exprimant les relations qui peuvent exister entre les entités spatiales (connexion, intersection, appartenance, etc.).

Nous avons analysé ces modèles dans le cadre particulier d'une expression textuelle de l'information spatiale. En effet, notre cas d'études était un corpus historique et territorial provenant d'une médiathèque, dont l'objectif était sa valorisation via des propositions innovantes de parcours et de recherche. Les

modèles cités ci-dessus nous ont permis de bâtir des méthodes d'extraction et d'interprétation de l'information. L'hypothèse de départ de notre travail consiste donc à considérer l'information spatiale comme étant une entité composée d'entité(s) nommée(s) et d'indicateurs spatiaux. Nous proposons un modèle bâti sur cette idée et verrons comment il est utilisé dans les méthodes d'extraction présentes dans notre plate-forme de recherche d'information spatiale.

Cependant, il faut noter au préalable que nous considérons l'information spatiale comme étant définie à l'intérieur d'une information plus complète, l'information géographique. Nous considérons en effet cette dernière (Gaio 2001) comme étant une molécule (fig. 1) composée des trois atomes « entité spatiale », « entité temporelle » et « phénomène ». Notre travail a donc consisté à travailler sur la partie spatiale tout en gardant à l'esprit le fait qu'elle fait partie de la composante englobante « entité géographique ».

Notre réalisation a d'abord consisté en une plateforme de recherche d'information spatiale dans du texte. Une deuxième partie reprend le concept de molécule géographique et se place à une échelle supérieure de représentation de l'information correspondant à une agrégation de ces molécules. Elle utilise les résultats de la première partie pour agréger l'information spatiale à un niveau plus élevé d'abstraction. Une illustration possible des résultats attendus est l'agrégation d'entités spatiales dans un paragraphe décrivant un itinéraire. Cet itinéraire correspond à une information spatiale plus précise et de portée plus importante. Ce travail permet d'imaginer un outil d'indexation et de recherche multi-niveaux de l'information géographique.

2 Modèle de l'entité géographique

Nous considérons donc que l'entité spatiale est à définir au sein d'une entité géographique. C'est pourquoi nous reprenons la définition de molécule géographique, concept déjà présent dans plusieurs travaux (Gaio 2001 ; Usery et al. 2004). L'information géographique peut être définie comme une molécule formée d'une composante spatiale, d'une composante temporelle et d'une composante thématique ou phénomène (fig. 1) (Malandain 2003 ; Perry et al. 2007). Cette définition vient du monde des bases de données. Nous l'avons néanmoins retenue en faisant l'hypothèse que le même type d'information peut être retrouvé et extrait de données non structurées (documents textuels).

Dans le cadre de nos travaux de recherche, nous avons proposé néanmoins une définition plus restreinte : les entités géographiques (EG), auxquelles nous allons nous intéresser, possèdent forcément une composante spatiale (ES) explicite. Celle-ci consiste en une ou plusieurs entités nommées de lieux *(*ville de Pau, les pics de la chaîne de la Maladetta*).* Par contre, l'entité temporelle (ET), qui peut être une date, un intervalle de dates ou une période *(*XVIII^{e} siècle, le 12 juin 1876, le début des années 60*)*, peut être implicite, c'est-à-dire qu'elle n'est pas mentionnée directement dans le texte mais découle d'informations annexes. L'ET peut être éloignée des autres composantes formant l'EG (par exemple dans le cas d'un journal de bord, la date est indiquée en début de paragraphe et le reste de l'unité de texte décrit des phénomènes se passant à cette date). L'ET peut donc aussi être associée à plusieurs EG, quand elle recouvre un paragraphe entier par exemple. Enfin le phénomène, ou composante thématique, correspond a priori à tout ce qui n'est pas spatial ou temporel dans le texte. C'est le sujet dont il est question à un lieu et moment donnés (botanique, thermes, etc.). Cependant, sa présence n'est pas obligatoire dans la molécule géographique dans la mesure où l'on considère que le thème prédominant peut être le phénomène géographique lui-même. En effet, il peut arriver que le sujet en question soit l'itinéraire pris par l'auteur. La description de cet itinéraire peut alors être considérée comme un phénomène. La molécule géographique est formée dans ce cas par la représentation de l'itinéraire dans l'espace, dans le temps et dans la manière dont il a été réalisé (Lesbegueries et al. 2006).

3 Modèle de l'entité spatiale (modèle pivot)

Nous définissons ici l'entité spatiale telle qu'elle peut être présente dans la molécule géographique décrite ci-dessus. D'après l'hypothèse de cible - site de Vandeloise, l'information spatiale exprimée dans un texte est constituée d'au moins une entité nommée et d'un nombre variable d'indicateurs spatiaux, précisant sa localisation. Nous sommes partis de ce principe pour définir un modèle cognitif pour l'information spatiale.

Ce modèle, nommé pivot car il servira pour toutes les étapes d'indexation et de recherche d'information, comporte quelques similitudes avec les modèles existants, comme celui décrivant l'ontologie du projet SPIRIT (Jones 2002). Une entité spatiale définie dans notre modèle peut donc correspondre à l'une des deux options suivantes (fig. 2) :

- Une entité spatiale absolue (ESA), dans le cas où l'auteur exprime une information seulement à partir d'une entité nommée (ex : la ville de Pau, Laruns*)*. Ce sont, en quelque sorte, les primitives spatiales de notre modèle.

- Une entité spatiale relative (ESR), dans le cas où l'auteur utilise en plus des entités nommées, des indications spatiales d'ordre topologique *(*ex : près de Pau, au sud de Pau, à une heure de marche de Pau, entre Pau et Laruns*)*.

Nous appelons ces indications spatiales des relations. En nous basant sur les modèles de raisonnement spatial qualitatif existants (Cohn 1997), nous en avons défini cinq :

- l'orientation (au sud de),
- la distance (à 1 heure de marche de, à 20 km de),
- l'adjacence (près de, loin de, la périphérie de),
- l'inclusion (le quartier de, la frontière entre, le sommet de),
- l'union et l'intersection liant au moins 2 ES (entre A et B, le triangle A, B, C, à l'intersection de A et B, la frontière A-B, etc.).

Une ES est donc définie par une de ces relations et au moins une autre ES (il peut y en avoir plusieurs dans le cas des relations d'union et d'intersection). Cette définition peut être récursive si la dernière ES est à son tour une ESR. Cette notion de définition récursive est un atout de notre modèle du point de vue pragmatique mais aussi du point de vue cognitif, car l'interprétation qui en découle reste proche de l'expression de la spatialité dans du texte. À titre d'exemple, quand un auteur utilise une ES complexe, du type « au nord de la frontière franco-espagnole », celle-ci est une ESR composée d'une relation d'orientation (« au nord de ») et d'une autre ESR (« frontière franco-espagnole »), elle-même composée d'une relation d'inclusion (« la frontière») et des ESA « France » et « Espagne ».

Le modèle ainsi présenté est capable d'interpréter la plupart des informations spatiales exprimées en langage naturel, en utilisant toutes les relations qualitatives possibles. Nous associons ensuite à chaque ES au moins une représentation. Cette représentation est géo-référencée et il peut y en avoir plusieurs, suivant les différentes échelles à laquelle est observée l'ES. En effet, une ville peut être représentée sous forme de point à l'échelle du pays, mais si le cadre d'étude est de l'ordre de la région, sa représentation correspondra plus à une forme polygonale. L'indexation ainsi construite dans le cadre d'un processus de recherche d'information est donc formée, pour chaque entité spatiale interprétée, d'une structure correspondant à une instance du modèle pivot et d'au moins un représentation géo-référencée.

De la même manière, le modèle pivot est utilisé pour l'interrogation, dans la mesure où les requêtes de l'utilisateur sont interprétées de la même façon que les documents. La requête, qui peut être exprimée en texte libre ou en dessinant, par exemple, des formes géométriques sur une carte, produit à l'aide du modèle pivot une ou plusieurs représentations géo-référencées qui sont mises en correspondance avec celles indexées dans les documents. L'appariement réalise alors un calcul sur les surfaces d'intersection afin de fournir un score de pertinence. De plus, le fait d'utiliser le même processus d'indexation pour les documents et les requêtes permet d'imaginer une interrogation du système via un document-requête, à condition de formaliser les opérateurs logiques (et, ou) entre les entités extraites.

4 Extraction d'information spatiale dans le texte

Des divers travaux existant sur l'analyse linguistique (Bilhaut 2006 ; Abolhassani 2003) se dégage un processus standardisé : segmentation, analyse lexicale et morphologique, analyse syntaxique, analyse « sémantique » liée à un domaine d'étude. Nous nous sommes basés sur ces travaux et sur le modèle pivot pour construire l'extraction d'information spatiale (fig. 3). La particularité du traitement réside au niveau de l'analyse morpho-syntaxique, qui extrait les entités nommées, et au niveau de l'analyse « sémantique », implémentées par une grammaire récupérant les indicateurs se trouvant aux alentours de ces entités. Le processus produit alors des instances du modèle pivot (entités spatiales absolues ou relatives stockées dans un format XML). Un module dédié prend alors en entrée les instances du modèle pivot et produit pour chacune d'entre elles des empreintes géo-référencées. Ces empreintes constituent alors un index spatial qu'un module d'appariement (utilisant des calculs d'intersection) parcourt pour renvoyer les entités pertinentes par rapport à une requête spatiale.

5 Plate-forme de recherche d'information spatiale

À partir des modèles définis, une plate-forme est proposée, constituée par de multiples briques de traitement unitaire, allant de l'extraction par analyse linguistique au géo-référencement et à la recherche

des entités spatiales trouvées dans des documents textuels. La figure 4 représente le schéma fonctionnel de cette plate-forme. La partie haute décrit le processus d'indexation, au cours duquel le corpus numérisé et OCR-isé subit un traitement linguistique. Les résultats de ce traitement sont des instances du modèle pivot, stockées à part. Ensuite un module de validation et de géo-référencement produit des empreintes géo-référencées qui sont stockées sous forme d'index dans une base grâce à un module dédié. La partie basse correspond au processus de recherche d'information. Une requête en texte libre comprenant une entité spatiale est analysée par les mêmes modules que ceux utilisés pour l'indexation. Le résultat forme aussi une empreinte géo-référencée qui est mise en entrée du module d'appariement. Celui-ci utilise alors l'index pour retrouver des paragraphes spatialement pertinents (contenant des entités spatiales qui s'intersectent avec celle de la requête).

6 Classification d'unités de texte en contextes

Les premiers travaux implémentés dans le cadre du prototype PIV donnent une indexation spatiale à un premier niveau d'abstraction (intra-phrastique). L'information sur laquelle on peut effectuer des recherches consiste en des syntagmes nominaux se trouvant dans des paragraphes. Cependant, celle-ci est souvent utilisée dans un contexte particulier. Plusieurs syntagmes nominaux, à l'intérieur d'un paragraphe, forment par exemple la description d'un itinéraire (je suis parti de A, près de B, pour aller à C, etc.). Détecter les contextes dans lesquels sont employés ces syntagmes peut apporter une recherche d'information spatiale plus fine. Nous avons donc proposé une méthode de classification de contextes spatiaux permettant d'avoir une indexation à un plus haut niveau d'abstraction.

Pour cela nous avons proposé la méthode suivante :

- regroupement des entités spatiales (syntagmes spatiaux) par unité de texte (paragraphe) ;
- calcul de caractéristiques informant sur les propriétés de dispersion et de linéarité entre les entités spatiales ;
- méthode d'apprentissage supervisée (pour classer les unités de texte en divers contextes spatiaux) utilisant ces caractéristiques.

Une première expérimentation a été réalisée sur un petit jeu de test. 74 paragraphes ont été lus et classés manuellement, 5 classes ont été définies : itinéraire, description de point de vue, description locale (localisée), comparaison de lieux et autre (pour les paragraphes ne faisant pas partie des 4 premières classes). La classification obtenue atteint une précision de 50% (une chance sur 2 pour 5 classes).

7 Conclusion des travaux de thèse

Nous proposons dans cette thèse des méthodes d'extraction d'information spatiale dans un texte. Ces méthodes proposent d'abord un premier niveau d'extraction, dégageant du texte des syntagmes nominaux. Afin de valider cette première étape, un prototype d'indexation et de recherche d'information spatiale a été bâti sur le modèle pivot, modèle cognitif et opérationnel permettant de manipuler l'information spatiale. Une deuxième étape a consisté à extraire du texte, non plus des atomes d'information spatiale, mais des molécules entières véhiculant un contexte spatial particulier tel qu'un itinéraire, une comparaison de lieux, une description, etc. Ces méthodes permettent d'imaginer une agrégation de l'information spatiale (et a fortiori de l'information géographique) à plusieurs niveaux d'abstraction jusqu'à reconstruire un « résumé » pour une œuvre entière, dans le cadre d'un système de recherche d'information.

Les perspectives pour ces travaux peuvent être décomposées en deux parties. La première consiste à finaliser la procédure d'indexation multi-niveaux pour la composante spatiale. Il reste à valider et à généraliser le processus de classification en contextes. En effet, la classification réalisée pour l'instant n'abstrait l'information spatiale qu'à un premier niveau. Il faudra imaginer des mécanismes d'agrégation entre ces premiers « résumés spatiaux » pour indexer le document texte à des niveaux de plus en plus élevés jusqu'au document lui même. La deuxième partie consiste à compléter la gestion de la molécule géographique, en proposant un modèle pour la partie temporelle, similaire au modèle proposé pour la partie spatiale, ainsi qu'à gérer la partie thématique. Il sera alors possible de représenter un document à partir de molécules décrivant tout ou partie de celui-ci (paragraphe, chapitre, etc.). Une plate-forme de système de recherche d'information spécialisée pourrait alors répondre à des requêtes complexes mettant en jeu des sous-requêtes spatiales, temporelles et thématiques.

Bibliographie

Abolhassani M., Fuhr N. et Gövert N., 2003, « Information extraction and automatic markup for xml documents, Intelligence Search on XML data », *LNCS Springer,* p. 159–178.

Bilhaut F., 2006, *Analyse automatique de structures thématiques discursives - Application à la recherche d'information*, mémoire de thèse, Université de Caen.

Cohn A. G., 1997, "Qualitative spatial representation and reasoning techniques", dans *KI '97 : Proceedings of the 21st Annual German Conference on Artificial Intelligence*, London, UK, Springer-Verlag, p. 1–30.

Denis M., 1997, *Langage et cognition spatiale*, Paris, Masson.

Egenhofer M., 1991, "Reasoning about binary topological relations, Second Symposium on Large Spatial Databases, Zurich, Switzerland", *Lecture Notes in Computer Science*, 525, p.143–160.

Gaio M., 2001, *Traitements de l'information géographique : représentations et structures,* mémoire d'HDR, Université de Caen.

Jones Chr. B., Purves R., Ruas A., Sanderson M., Sester M., Van Kreveld M. et Weibel R., 2002, "Spatial information retrieval and geographical ontologies : an overview of the SPIRIT project", dans *25th ACM SIGIR*.

Lesbegueries J., Gaio M. et Loustau P., 2006, "Geographical information access for non-structured data", dans *ACM Symposium on Applied Computing (SAC), Dijon, France*, p. 83–89.

Malandain N., 2003, *La relation Texte/Image, essai de modélisation dans un corpus géographique*, mémoire de thèse, Université de Caen.

Perry M., Sheth A. et Arpinar I. B., 2007, "Geospatial and Temporal Semantic Analytics", dans *Encyclopedia of Geoinformatics,* Hassan A. Karimi (Ed), Idea-Group, p. 1–14.

Usery E. L., Timson G. et Coletti M., 2004, "Multidimensional representation of geographic features", *International Journal of Geographic Information Science,* p. 1–8.

Vandeloise Cl., 1986, *L'espace en français,* Paris, Éditions du Seuil.

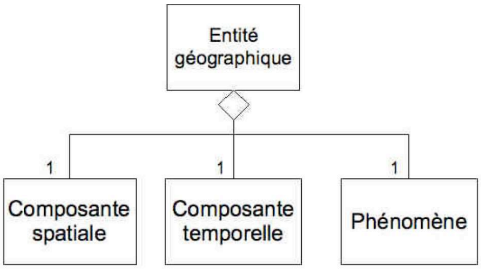

Figure 1 : Définition de l'entité géographique sous forme de molécule

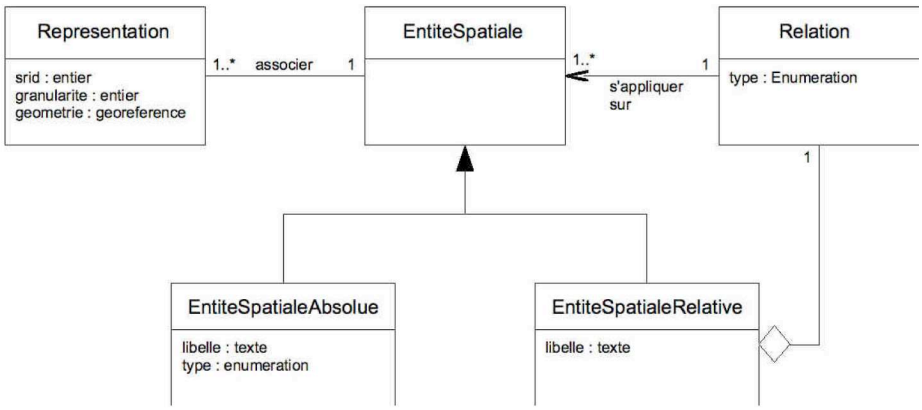

Figure 2 : Modèle spatial pivot

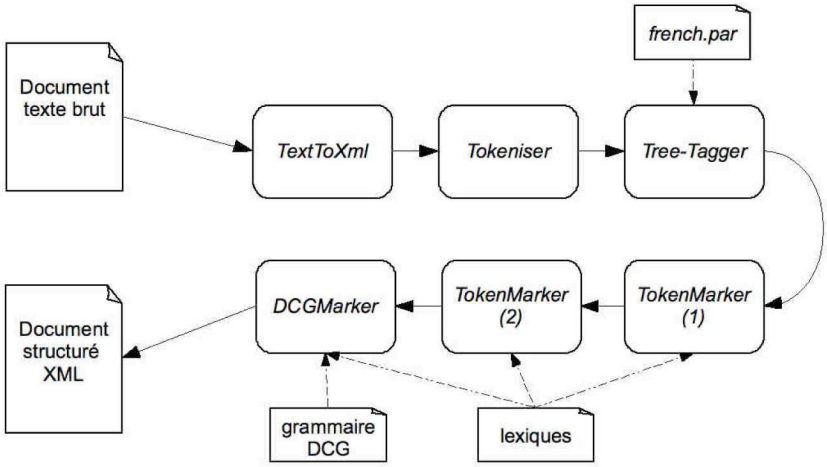

Figure 3 : Chaîne de traitement pour l'indexation spatiale

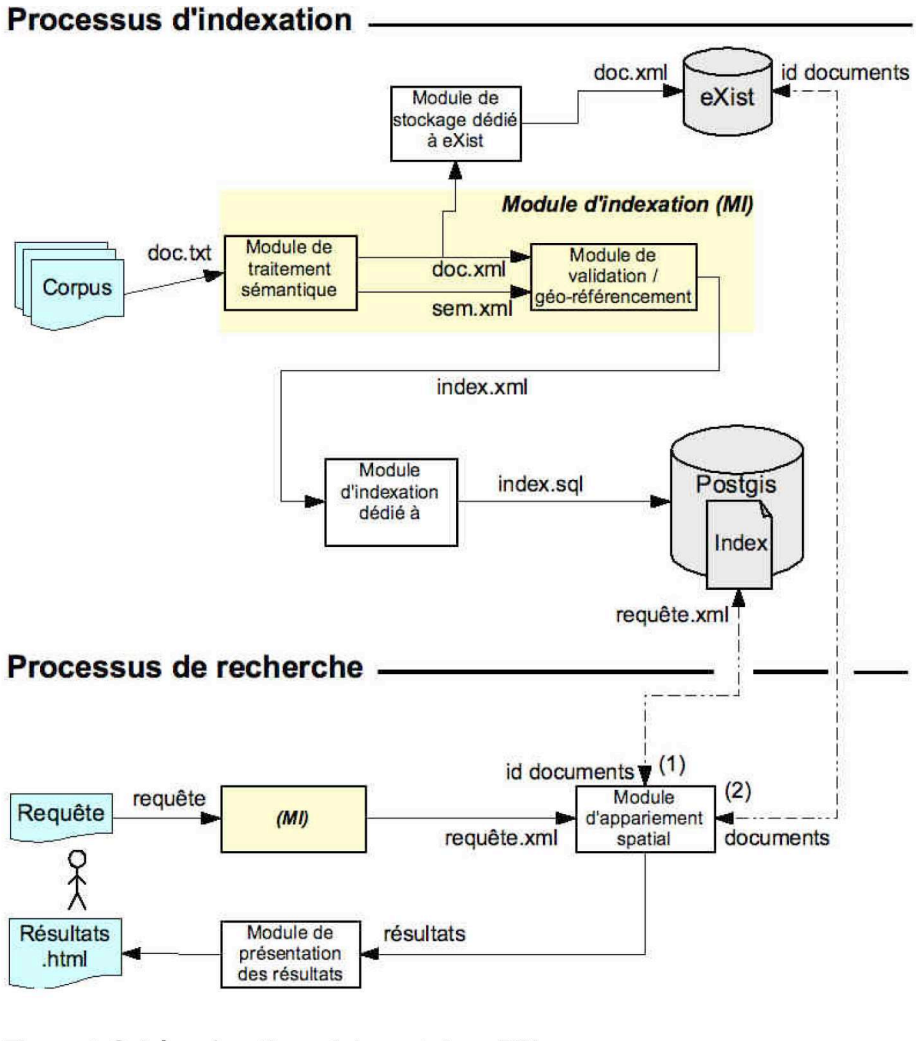

Figure 4 : Schéma fonctionnel du prototype PIV