La composante temps dans l'information géographique textuelle

Extraction et recherche d'information dans des fonds documentaires patrimoniaux numérisés

Annig Le Parc - Lacayrelle — Mauro Gaio — Christian Sallaberry

LIUPPA - EA 3000

Université de Pau et des Pays de l'Adour

64013 Pau cedex, France

{annig.lacayrelle, mauro.gaio, christian.sallaberry}@univ-pau.fr

RÉSUMÉ. Les travaux présentés dans ce papier se situent dans le cadre de la recherche d'information géographique. Notre objectif est d'exploiter les fonds documentaires patrimoniaux dont le contenu est fortement attaché au territoire (information géographique). L'information géographique peut être déclinée selon le triptyque espace, temps, phénomène. Ce papier se focalise plus particulièrement sur le volet temps. Nous proposons tout d'abord un modèle décrivant les expressions temporelles à annoter. Puis, nous présentons une méthodologie d'annotation sémantique pour l'indexation automatique de ces expressions. Enfin, nous évoquerons la recherche d'information géographique d'abord selon les critères spatial, thématique et temporel respectivement, puis nous combinons ces trois critères.

ABSTRACT. These works take place in the geographical information retrieval context. Our purpose is to exploit local cultural heritage document repositories whose content is strongly attached to a territory (i.e. geographical information). Geographical information is composed of a spatial feature, a temporal feature and a theme feature. This paper focuses on temporal feature. We first propose a model for temporal expressions to annotate. Then, we present a semantic annotation methodology for automatically indexing these expressions. Finally, we start on geographical information retrieval according respectively to spatial, theme and temporal criterions, and then we combine them.

MOTS-CLÉS : recherche d'information, information spatiale, information temporelle, information thématique, sémantique des contenus

KEYWORDS: information retrieval, spatial information, temporal information, theme information, contents' semantics

1. Introduction

Aujourd'hui, de nombreuses institutions comme les archives départementales ou régionales, les bibliothèques et médiathèques ou les musées ont opté pour la numérisation de leurs fonds documentaires. La numérisation de ces fonds permet de proposer de nouvelles formes de valorisation des documents grâce à l'usage des nouvelles technologies, un processus de valorisation se traduisant par la mise en place de moyens spécifiques pour véhiculer, de manière ciblée, un contenu sur un public donné en vue d'un impact précis.

En outre, une étude faite sur le moteur de recherche Excite montre qu'une requête sur cinq effectuée sur celui-ci est liée à la géographie (Sanderson *et al.*, 2004).

L'information géographique, contenue dans le corpus mis à notre disposition par la MIDR¹, est composée d'entités spatiales, temporelles et thématiques. « Les instruments de musique dans le voisinage de Laruns au XIX^e siècle » est un exemple d'entité géographique complète : « Instruments de musique » est l'entité thématique, « dans le voisinage de Laruns » est l'entité spatiale et « au XIX^e siècle » l'entité temporelle.

La plupart des systèmes permettant la gestion et la consultation de documents numériques en ligne proposent des méthodes d'indexation reposant sur l'exploitation de méta-données², combinées à des techniques de recherche plein texte. Ces dernières sont basées essentiellement sur des méthodes d'analyse de données portant sur les formes (mots ou graphies) qui constituent un texte. Seules quelques propositions pré-industrielles améliorent la précision grâce à la prise en compte de certaines relations sémantiques dans la requête de l'utilisateur (comme, par exemple, certaines relations spatiales dans le projet européen SPIRIT (Chen *et al.*, 2006)). Mais dans ces propositions, l'indexation géographique des contenus, quant à elle, se limite à associer une entité nommée toponomique à une géoréférence.

Actuellement donc, les techniques d'indexation opérationnelles ne permettent pas une prise en compte aisée de l'information géographique telle qu'elle apparaît dans notre corpus. Or parallèlement et depuis bientôt une dizaine d'années, l'analyse sémantique de contenus à partir de textes connaît une évolution importante grâce notamment à un renouvellement de la manière de poser le problème (Bourigault *et al.*, 2001) (Enjalbert *et al.*, 2004). Les textes sont vus désormais comme des sources à exploiter par des logiciels de Traitement Automatique des Langues (TAL) au sein d'un processus supervisé par un analyste qui tient compte des objectifs de la modélisation.

Aux méthodes de recherche basées sur l'analyse de données, de plus en plus de travaux opposent donc actuellement des méthodes permettant d'exploiter véritablement la sémantique du « contenu » des documents. Celle-ci est appréhendée de manière très partielle pour des raisons de complexité. Le gain attendu est à la fois en termes de rap-

1. Médiathèque Intercommunale à Dimension Régionale.

2. Nous entendons par méta-données un ensemble de descripteurs synthétisant l'ensemble du contenu d'un document et produit semi-automatiquement voire manuellement.

pel, de précision (car cela permet de dépasser la combinaison booléenne d'indicateurs) et de meilleure compréhension par l'utilisateur des résultats de sa requête.

Dans le domaine de l'information géographique, l'enjeu est donc de pouvoir « interpréter » des expressions telles que « *au Nord de* », « *entre* », « *à proximité de* » ou « *en périphérie de* » ou temporelles comme « *vers le début* », « *aux alentours de* » voire de composition ou de dépendance « *les pics Pyrénéens les plus élevés* » ou « *les gaves à fort débit* » ou encore « *vers le début de l'hiver 1850* » ou alors « *après les années de sécheresse du début de ce siècle* ». Deux approches majeures sont aujourd'hui proposées afin de prendre en compte la sémantique du contenu du document. La première approche porte sur des techniques basées sur des ontologies de domaine. Avec ce type de technique, la structure de l'index est donnée par la structure de l'ontologie employée. Un des problèmes rencontrés vient de la difficulté de disposer d'ontologies offrant des concepts spécifiquement dédiés à l'information géographique. En effet, très peu d'ontologies sont actuellement réalisées en pratique (Lemmens, 2006) ou alors celles-ci portent sur des domaines très ciblés. La seconde s'appuie sur des techniques TAL (traitements automatiques de la langue) permettant de construire une représentation symbolique d'une information cible présente dans le contenu des documents. Cela consiste généralement à appliquer au flux textuel des patrons syntaxico-sémantiques associés à des bases lexicales. Un des problèmes rencontrés vient du fait d'une grande variation des performances de ces techniques celles-ci dépendant de deux paramètres : la complexité/variabilité de l'information recherchée et le corpus sur lequel les patrons ont été définis (Aussenac-Gilles *et al.*, 2006).

Dans le cadre du projet PIV³, à l'origine des travaux présentés dans cet article, la seconde approche est bien adaptée. En effet, si l'on décline l'information géographique selon le triptyque espace, temps, thème/phénomène et que l'on opère, uniquement et de manière indépendante, sur les composantes espace et temps (connues comme bénéficiant d'une certaine stabilité dans la langue), on peut considérer que le problème posé par la complexité/variabilité devient maîtrisable.

Enfin, l'information géographique dans des corpus textuels peut être exprimée dans des formes plus ou moins explicites (complètes). L'implicite (incomplète) peut se situer au niveau des différentes composantes (espace, temps, thème/phénomène) mais également à l'intérieur de chacune d'entre elles. Cette propriété a orienté notre démarche aussi bien, dans la manière de modéliser les deux composantes (espace et temps) sur lesquelles portent nos traitements sémantiques, que dans la manière de construire les index.

La section 2 présente, dans le cadre de la démarche de modélisation et de marquage des expressions textuelles, nos propositions relatives à la composante temporelle. La section 3 quant à elle décrit, toujours en ce qui concerne la composante temporelle de l'information géographique, l'approche retenue pour produire une interprétation formelle et les index correspondants. La dernière section détaille dans un premier

3. « Pyrénées Itinéraires Virtuels » : projet cofinancé par la Communauté d'Agglomération de Pau et en partenariat avec la MIDR.

temps les différentes approches adoptées pour chaque composante de l'information géographique (espace, temps, phénomène/thème) afin d'élaborer un système de recherche d'information (RI) par le contenu. Cela permettant d'aboutir à un système de recherche s'appuyant sur une granularité intra-documentaire de l'information. Enfin, dans un second temps, nous proposons une première technique permettant de combiner les résultats obtenus avec les recherches spatiale, temporelle et thématique.

2. Le temps dans les corpus territorialisés

Dans cette section, nous allons détailler les expressions temporelles typiques que l'on trouve dans notre corpus⁴. Nous étudierons les travaux réalisés pour la modélisation du temps et notamment le langage timeML. Enfin, nous présenterons notre modèle.

2.1. Expressions temporelles de notre corpus

Notre objectif étant de connaître les périodes abordées dans les documents, nous nous intéressons seulement aux expressions temporelles à connotation calendaire. Voici des exemples d'expressions temporelles que l'on peut trouver dans nos documents :

11 mars 1957 ; juillet 1957 ; 11 mars ; mars ; printemps 1787 ; été de 1799 ; hiver ; 2002 ; XVIII^e siècle ; 18h00 ; fin de 1892 ; dès 1781 ; depuis 1787 ; 1997-1998 ; fin août 1801 ; de 1950 à 1960 ; début des années 1950 ; vers la fin du printemps de 1800 ; avril et mai 1798 ; premiers jours de l'automne ; vers la fin de 1794 ou le début de 1795 ; 12, 15 et 20 juillet 1916 ; 1950,52,57 ; du 29 au 30 septembre ; avant l'ère chrétienne ; depuis la révolution ; pendant 18 ans.

Dans un premier temps, nous avons décidé de ne pas retenir les expressions faisant référence à des heures (car dans notre fonds documentaire, ce n'est pas un grain approprié), à des durées (à cause de leur référence anaphorique, il est difficile de les placer sur un axe du temps), et à des événements historiques (car elles nécessitent un thésaurus d'événements historiques). Dans le sous-ensemble restant, certaines expressions référencent explicitement une entrée dans le calendrier : ce sont des dates (ex : 11 mars 1957 ; printemps 1901 ; XVIII^e siècle ; ...). Nous les appellerons entités absolues. D'autres correspondent à une adaptation d'une ou plusieurs entités temporelles absolues (ex : fin de 1892 ; vers la fin du printemps de 1800 ; ...). Nous les appellerons entités temporelles relatives. La granularité la plus fine est le jour. Les entités temporelles peuvent être complètes ou incomplètes. Les entités incomplètes ne peuvent pas être situées sur une échelle de temps absolue car elles ne possèdent pas d'année (ex : mars ; 11 mars ; hiver ; ...).

4. Le corpus mis à la disposition par la MIDR est principalement constitué de monographies relatant des recits de voyages (principalement du XIX^e siècle) dans les Pyrénées.

2.2. Modélisation du temps

De nombreux travaux sont consacrés à l'analyse temporelle du document. (MUC-7, 1998) ne s'est intéressé qu'aux entités temporelles absolues (date et heure). De façon plus générale, (Vazov, 2001) identifie dans le document les entités qui répondent à la question « *quand* », mais ne leur associe ni sémantique ni représentation. (Schilder *et al.*, 2001) propose un marquage sémantique des entités temporelles calendaires, mais une représentation non uniforme. Enfin, des travaux tels que (Gosselin, 1996), (Mani *et al.*, 2000), (Setzer *et al.*, 2002), (Muller *et al.*, 2004), (Mani, 2004) et (Sauri *et al.*, 2006) s'intéressent à l'analyse temporelle de la langue naturelle en marquant les relations temporelles qui existent entre les événements contenus dans les documents. (Muller *et al.*, 2004) s'intéresse plus particulièrement à l'annotation automatique des relations introduites par des verbes conjugués. Les travaux tels que ceux de (Mani *et al.*, 2000), (Wilson *et al.*, 2001), (Setzer *et al.*, 2002) et (Ferro *et al.*, 2003) ont conduit à l'unification de la manière d'annoter le temps dans les documents (expressions calendaires et relations entre événements) au travers du langage timeML (Sauri *et al.*, 2006).

TimeML utilise quatre types de balises, chacune possédant un ensemble d'attributs permettant de décrire précisément l'information marquée. Les différentes balises sont décrites dans le tableau 1.

<EVENT>	permet de marquer les événements, ponctuels ou ayant une certaine durée, qui se produisent ou qui se sont produits
<TIMEX3>	permet de marquer les heures, les dates, les intervalles et les durées. Ils peuvent être complètement spécifiés (ex : « le 3 juin 1950 ») ou non (ex : « le 3 juin »)
<SIGNAL>	permet d'annoter les mots indiquant comment les informations temporelles sont reliées (ex : avant, après, durant, ...)
<TLINK>	Temporal link : permet de décrire les relations temporelles entre les balises <EVENT>, entre les balises <TIMEX3>, ou entre les balises <EVENT> et <TIMEX3>
<ALINK>	Aspectual link : permet de décrire les relations entre les balises <EVENT> correspondant à un événement aspectuel (commencer, finir, ...) et les balises <EVENT> correspondant à l'événement argument
<SLINK>	Subordination link : permet de décrire le contexte introduisant les relations entre deux balises <EVENT>

Tableau 1. Les balises timeML

Etant donné que nous nous intéressons seulement aux entités de type calendaire, seules les balises <TIMEX3>, <SIGNAL> et <TLINK> ont un intérêt pour le marquage de nos documents.

Prenons quelques exemples :

3 juin 1963	<code><TIMEX3 tid="t1" type="DATE" value="1963-06-03"> 3 juin 1963 </TIMEX3></code>
été 1963	<code><TIMEX3 tid="t2" type="DATE" value="1963-SU"> été 1963 </TIMEX3></code>
juin 1963	<code><TIMEX3 tid="t3" type="DATE" value='1963-06'> juin 1963 </TIMEX3></code>

L'attribut `value` utilise la notation préconisée par la norme ISO-8601. La forme de la valeur dépend de la nature de l'entité marquée. Notre objectif étant de marquer des entités temporelles pour un usage plus efficace en RI (Recherche d'Information), cette forme de représentation présente des limites : en effet, lors de la phase de recherche, les comparaisons vont être réalisées entre éléments ayant des formes différentes.

Lorsque l'entité calendaire est relative comme, par exemple, « *le début de juin 1963* », la marquage se fait de la façon suivante :

```
<SIGNAL sid="s1">début</SIGNAL>
```

de

```
<TIMEX3 tid="t1" type="DATE" value="1963-06">juin 1963 </TIMEX3>
```

La construction de « *début de juin 1963* » ne peut se faire qu'en la rattachant avec l'événement associé. Soit *ei4*, l'identifiant de l'instance de cet événement. On obtient alors le marquage suivant :

```
<TLINK eventInstanceID="ei4" relatedToTime="t1" signalID="s1"
relType="BEGINS" />
```

L'attribut `relType` précise le type de la relation temporelle. Le problème est qu'aucune représentation (valeur) n'est associée à « *début de juin 1963* ». Or, lors de la phase de recherche, il est nécessaire d'avoir une représentation pour cette période.

Notre objectif est de permettre la recherche de documents traitant d'une période donnée, sachant que la granularité la plus fine est le jour. L'annotation des éléments temporels contenus dans les documents devra être réalisée de manière automatique. D'une part, le fait que les entités temporelles relatives ne sont pas associées à des représentations, et d'autre part le fait que la représentation des entités temporelles absolues n'est pas uniforme constituent un inconvénient pour la recherche. Ce sont pour ces raisons que nous n'avons pas utilisé le langage timeML pour annoter nos documents. Nous avons donc proposé un langage pour l'annotation conforme à notre modèle temporel, ce dernier étant en adéquation avec notre modèle spatial (Gaio *et al.*, à paraître en décembre 2007).

2.3. Notre proposition

L'élaboration du modèle a commencé par l'étude de notre corpus. Suite à cette étude, nous avons proposé un modèle unifié pour décrire les entités spatiales et temporelles. La figure 1 donne un aperçu du modèle.

Figure 1. Schéma simplifié des modèles spatial et temporel

Les entités spatiales (ES) comme les entités temporelles (ET) peuvent être absolues ou relatives. Une entité spatiale absolue (ES_A) correspond à une entité géolocalisée. Une entité spatiale relative (ES_R) définit une relation spatiale avec au moins une ES. Par exemple, « au nord de Pau » est une ES_R qui définit une relation d'orientation avec l' ES_A « Pau ». Pour plus de détails sur le modèle spatial, voir (Lesbeguerries *et al.*, 2006).

Une entité temporelle absolue (ET_A) correspond à une date. Une entité temporelle relative (ET_R) décrit une relation entre une ou plusieurs ET (qui peuvent être absolues ou relatives). Chaque ET possède une ou plusieurs représentations. Nous avons choisi de représenter les ET à l'aide d'intervalle de temps (comme dans le modèle SdT (Gosselin, 1996)). L'intérêt est de nous permettre, lors de la phase de recherche, de ne calculer que des intersections d'intervalle et notamment d'utiliser les opérateurs définis par Allen (Allen, 1984).

Les ET_A correspondent à des dates : un jour, un mois, une saison, une année ou un siècle. Elles peuvent être complètes ou incomplètes. Par incomplet, on entend un jour, un mois ou une saison sans année.

Les ET_R correspondent à une relation entre une ou plusieurs ET. Nous distinguons cinq types de relation :

- inclusion : relation unaire qui permet de décrire des entités temporelles telles que « au début de ET », « à la fin de ET », ...
- adjacence : c'est une relation unaire qui permet de décrire des entités temporelles telles que « vers ET », « aux alentours de ET », « les années ET », ...
- orientation : c'est une relation unaire qui permet de décrire des entités temporelles telles que « avant ET », « après ET », ...
- intervalle : c'est une relation binaire qui permet de décrire des entités temporelles telles que « de ET_1 à ET_2 », ...

– énumération : c'est une relation n-aire qui permet de décrire des entités temporelles telles que « 9, 10 et 12 novembre 1900 », « 1950,60 », ...

Contrairement à timeML, notre modèle permet de décrire la sémantique d'une ET_R indépendamment de l'événement auquel elle est liée. De plus, le modèle associe une représentation unifiée à tout type d'ET sous la forme d'un ou de plusieurs intervalles de temps.

3. Extraction/Indexation des informations temporelles

La création des index temporels est réalisée en deux étapes. La première consiste à extraire les ET (décrites par notre modèle) à l'aide d'une chaîne de traitement syntaxico-sémantique. La deuxième consiste à interpréter la sémantique associée à l'ET pour calculer l'intervalle de temps associé.

3.1. Chaîne de traitement syntaxico-sémantique

La chaîne de traitement syntaxico-sémantique est réalisée à l'aide de la plateforme linguastream (Bilhaut, 2003) (Widlocher *et al.*, 2005). Elle prend en entrée un fichier au format txt et est composée principalement des traitements suivants :

- une analyse morpho-syntaxique à l'aide de l'outil Tree-Tagger (Schmid, 1994) ;
- une analyse lexicale à l'aide de l'outil Lexicon Marker ;
- une analyse syntaxico-sémantique réalisée à l'aide d'une grammaire DCG (Definite Clause Grammar) dans laquelle sont intégrés des traitements en Prolog.

Le résultat est un fichier xml contenant pour chaque ET marquée sa sémantique. L'exemple qui suit montre le balisage de l'ET « juillet 1957 ». Cette dernière correspond à une ET_A . La sémantique associée à chaque ET_A consiste en un grain, un début et une fin. Enfin, chaque ET possède un identifiant `et_id` et se situe dans le paragraphe `par_id`.

```
<et et_id="2" par_id="2" type_et="et_a">
  <texte> juillet 1957 </texte>
  <et_a>
    <grain> mois </grain>
    <debut jour="01" mois="07" annee="1957" />
    <fin jour="31" mois="07" annee="1957" />
  </et_a>
</et>
```

L'exemple suivant montre le balisage de l'ET « au début des années 1950 ». Cette ET est une ET_R qui correspond à une relation d'inclusion avec l' ET_R « années 1950 ». Cette dernière est une relation d'adjacence avec l' ET_A « 1950 ».


```

<et et_id="5" par_id="3" type_et="et_r">
  <texte> au début des années 1950 </texte>
  <et_r>
    <relation type="inclusion" sens="debut" />
    <et_r>
      <relation type="adjacence" sens="annees" />
      <et_a>
        <grain> annee </grain>
        <debut jour="01" mois="01" annee="1950" />
        <fin jour="31" mois="12" annee="1950" />
      </et_a>
    </et_r>
  </et_r>
</et>

```

3.2. Indexation

La phase d'indexation consiste à interpréter la sémantique de l'ET pour calculer le ou les intervalles de temps qui lui sont associés. Ce sont ces intervalles de temps qui seront utilisés lors de la phase de recherche.

3.2.1. Cas des ET_A

L'intervalle de temps associé à l' ET_A est le même que celui décrit dans la sémantique. Si nous reprenons l'exemple « *juillet 1957* », l'intervalle de temps associé va du 1957-07-01 au 1957-07-31. Nous avons utilisé la norme ISO-8601 pour les bornes de l'intervalle. Le balisage associé à l'ET devient donc :

```

<et et_id="2" par_id="2" type_et="et_a">
  <texte> juillet 1957 </texte>
  <et_a>
    <grain> mois </grain>
    <debut jour="01" mois="07" annee="1957" />
    <fin jour="31" mois="07" annee="1957" />
  </et_a>
  <periode>
    <debut> 1957-07-01 </debut>
    <fin> 1957-07-31 </fin>
  </periode>
</et>

```

3.2.2. Cas des ET_R

L'intervalle de temps associé à l' ET_R est calculé à partir de celui de l'ET (ou des ET) associée(s) en tenant compte de la nature de la relation. Soient e_r une ET_R en relation $\mathcal{R}$ avec l'ET e , I_{e_r} l'intervalle de temps associé à e_r et I_e l'intervalle de temps associé à e . On distingue plusieurs cas de figure (voir figure 2) :

Figure 2. Calcul des intervalles de temps associés aux ET_R

– si $\mathcal{R}$ = inclusion : I_{e_r} est inclus dans I_e . Les bornes de I_{e_r} dépendent du type d'inclusion (début, milieu, fin) et de la granularité de I_e . Par exemple, pour une inclusion de type début, I_{e_r} correspond environ au premier tiers de I_e ;

– si $\mathcal{R}$ = adjacence : I_{e_r} inclut I_e . I_e est donc élargie au niveau de ses deux bornes. Cet élargissement dépend de la nature de l'adjacence (vers, les années) et du grain de I_e ;

– si $\mathcal{R}$ = orientation : I_{e_r} inclut I_e . I_e est donc élargie au niveau d'une de ses deux bornes. Cet élargissement dépend de la nature de l'orientation (avant, après) et du grain de I_e ;

Soient e_r une ET_R en relation de type intervalle avec l'ET e_1 et avec l'ET e_2 , I_{e_r} l'intervalle de temps associé à e_r , I_{e_1} l'intervalle de temps associé à e_1 et I_{e_2} l'intervalle de temps associé à e_2 . I_{e_r} a pour borne inférieure la borne inférieure de I_{e_1} , et pour borne supérieure, celle de I_{e_2} (voir figure 2).

Prenons l'exemple « de 1950 à 1960 », l'intervalle de temps associé va du 1950-01-01 au 1960-12-31. Le balisage associé à l'ET est donc :

```
<et et_id="5" par_id="3" type_et="et_r">
  <texte> de 1950 à 1960 </texte>
  <et_r>
    <relation type="intervalle" />
    <et_a>
```

```

    <grain> annee </grain>
    <debut jour="01" mois="01" annee="1950" />
    <fin jour="31" mois="12" annee="1950" />
  </et_a>
  <et_a>
    <grain> annee </grain>
    <debut jour="01" mois="01" annee="1960" />
    <fin jour="31" mois="12" annee="1960" />
  </et_a>
</et_r>
<periode>
  <debut> 1950-01-01 </debut>
  <fin> 1960-12-31 </fin>
</periode>
</et>

```

Soient e_r une ET_R en relation de type énumération entre l' ET_A e_a et la liste de valeur val , et I_{e_a} l'intervalle de temps associé à e_a . Dans ce cas, plusieurs intervalles de temps sont associés à e_r : I_{e_a} plus autant d'intervalles qu'il y a d'éléments dans val . Ces intervalles sont calculés à partir de I_{e_a} et de sa granularité (voir figure 2). Par exemple, « 12, 15 et 20 juillet 1916 » est une ET_R de type énumération entre l' ET_A « 20 juillet 1916 » et la liste de valeur contenant 12 et 15. Ainsi, les intervalles de temps représentant « 12, 15 et 20 juillet 1916 » sont : [1916-07-12,1916-07-12], [1916-07-15,1916-07-15], [1916-07-20,1916-07-20].

Le processus d'interprétation est récursif. En effet, si nous reprenons l'exemple « au début des années 1950 », cette ET_R dépend d'une autre ET_R . Il faut donc commencer par associer un intervalle de temps à cette dernière (« les années 1950 »). Nous sommes dans le cas d'une relation d'adjacence de type années. L'intervalle de temps associé va donc du 1950-01-01 au 1959-12-31. Puis, nous calculons l'intervalle de temps final en prenant le premier tiers (relation d'inclusion de type début). Nous obtenons :

```

<et et_id="5" par_id="3" type_et="et_r">
  <texte> au début des années 1950 </texte>
  <et_r>
    <relation type="inclusion" sens="debut" />
    <et_r>
      <relation type="adjacence" sens="annees" />
      <et_a>
        <grain> annee </grain>
        <debut jour="01" mois="01" annee="1950" />
        <fin jour="31" mois="12" annee="1950" />
      </et_a>
    </et_r>
  </et_r>
</et_r>
<periode>
  <debut> 1950-01-01 </debut>
  <fin> 1953-12-31 </fin>

```

</periode>
</et>

3.3. Une première évaluation du processus d'extraction

Le corpus mis à notre disposition par la MIDR comporte une dizaine d'ouvrages. Nous avons réalisé une première évaluation de notre processus d'extraction en considérant un bon tiers de quatre ouvrages, ce qui représente environ 540 entités temporelles. Le tableau 2 montre comment se répartissent ces entités.

ET _A	ET _R	heure	réf. événement	durée
69,6 %	21,6 %	2,2 %	0,9 %	5,5 %

Tableau 2. Répartition des différents types d'entités temporelles

Cette répartition conforte notre choix de considérer le jour comme la granularité la plus fine. Le tableau 3 donne les taux de rappel et de précision obtenus pour les entités que nous voulons extraire (ET_A et ET_R).

	ET _A	ET _R
taux de rappel	0,91	0,90
taux de précision	0,97	1

Tableau 3. Répartition des différents types d'entités temporelles

Ces résultats s'expliquent par le fait que la grammaire a été définie à partir de l'étude des quatre ouvrages ayant servi à l'évaluation. Il faudrait maintenant réaliser une évaluation sur l'intégralité des ouvrages de notre corpus pour vraiment tester notre chaîne de traitement.

4. Vers une recherche d'information géographique

L'information géographique se déclinant selon le triptyque espace, temps, phénomène/thème, différentes approches d'extraction et d'indexation ont été proposées. Par exemple, une page web peut-être référencée par une annotation géolocalisée, d'une part et, un index full-text, d'autre part (Vaid *et al.*, 2005). Selon le même principe, T. Sagara (Sagara *et al.*, 2004) utilise les pages jaunes pour générer des listes de mot-clés (activités/métiers) et annoter des pages web. Les projets SPIRIT, GeoSearch, GEO-IR, liés à la gestion d'informations géographiques sont présentés dans (Chen *et al.*, 2006). Tous ces systèmes choisissent généralement le phénomène/thème comme point d'entrée de l'information géographique et proposent, comme SPIRIT, des index combinant cette information (hôtel/restaurant) avec l'information spatiale, par exemple. Toute la

difficulté est, bien entendu, d'assurer l'existence d'une réelle relation sémantique entre le phénomène/thème et l'information spatiale ainsi liés.

Notre approche vise des documents de type et d'origine divers, peu structurés, décrivant des territoires, et ne cible aucun phénomène/thème particulier a priori. Dans un contexte de recherche d'information géographique, le critère d'entrée principal dans le document est l'espace et/ou le temps. Ces critères sont plus stables que le phénomène/thème relativement varié et généralement inconnu a priori. Ainsi, nous avons mis en oeuvre des chaînes de traitement dédiées à l'information spatiale (Lesbeguerries *et al.*, 2006) et temporelle (cf. §3) afin d'étendre des systèmes d'indexation classiques basés sur des méthodes statistiques. Les index qui en résultent sont donc complètement indépendants. Ils permettent des recherches sur un seul ou plusieurs critères. Dans ce dernier cas, la combinaison des résultats est encore un problème de recherche très ouvert.

La RI géographique que nous voulons mettre en oeuvre ne considère pas les documents comme un tout. Nous travaillons donc sur une unité documentaire (appelée fragment de document) qui est la phrase dans le processus d'indexation et le paragraphe dans le processus de recherche. Le paragraphe permet de replacer les phrases dans leur contexte. La méthode de stockage adoptée permet de passer aisément du paragraphe à la page, et de la page au document.

4.1. RI spatiale

La recherche spatiale commence par une extraction des informations spatiales exprimées dans la requête ; ce traitement est similaire à la démarche de détection d'Entité Spatiale (ES) dans le fonds documentaire. Elle est suivie d'un appariement entre les ESs de la requête et celles contenues dans les index. Cet appariement est basé sur le calcul d'intersections entre les zones géo-référencées (boîtes englobantes) correspondant aux ESs de la requête et des index (Sallaberry *et al.*, 2007b).

Figure 3. Calcul de pertinence d'un document sélectionné

Pour chaque requête (voir figure 3), le système évalue son degré de similarité avec les fragments de documents de la collection en calculant d'abord le $Df_{precision}$ qui est

égal au rapport de la surface d'intersection entre la requête ($I_{surface}$) et le fragment de document ($Df_{surface}$) : $Df_{precision} = \frac{I_{surface}}{Df_{surface}}$

Puis le rapport avec la requête ($Df_{significance}$) : $Df_{significance} = \frac{I_{surface}}{Q_{surface}}$

Et le rapport de distance : $Df_{distance} = \frac{d}{D}$

Ainsi, le degré de similarité Df_{score} est calculée comme suit :

$$Df_{score} = \frac{(Df_{precision} + Df_{significance})}{(2 + Df_{distance})} \quad [1]$$

Plus les centroïdes de I et Q sont proches, plus la pertinence du fragment de document est grande.

4.2. RI thématique

La RI classique est basée sur la notion de « sac de mot » (Baeza-Yates *et al.*, 1999). Elle consiste d'abord à sélectionner les mots dans les documents, puis à les lemmatiser (Porter, 1980) et, enfin, à enlever les mots vides (stoplist). Nous avons utilisé une liste de mots vides et un lemmatiseur de la langue Française de la famille Snowball1. Un poids $Wtd(t, d)$ est ensuite assigné à chaque terme t d'un document d_j suivant la formule :

$$Wtd(t_j, d_j) = \frac{\frac{2tf_{ij} \times \log(N - n_i + 0,5)}{(n_i + 0,5)}}{2 \times (0,25 + 0,75 \times dl_j / avg_dl) + tf_{ij}} \quad [2]$$

$tf_{i,j}$ représente la fréquence du terme t_i dans le fragment de document d_j , n_i est le nombre de fragments contenant le terme t_i et N le nombre total de fragments dans la collection. dl_j représente la taille du fragment d_j et avg_dl , la taille moyenne de document dans la collection. Cette méthode de pondération, qui est une amélioration de la formule TF.IDF est introduite pour atténuer l'impact négatif des documents longs lors de la phase de recherche (Robertson *et al.*, 1998). Ceci est bien adapté aux paragraphes de tailles variées de notre collection. Le même processus d'indexation est appliqué aux requêtes. Un modèle de recherche vectoriel (Boughanem *et al.*, 2001) est ensuite utilisé pour la phase de recherche : pour une requête q donnée, le produit scalaire (Inner product) entre le vecteur de la requête et ceux correspondants à chaque fragment de document d_j dans la collection est appliqué pour calculer les scores de pertinence $Rel(q, d_j)$:

$$Rel(q, d_j) = \sum_{k=1}^{|q|} Wtq(t_k, q) \times Wtd(t_k, d) \quad [3]$$

Ce score de pertinence est utilisé pour déterminer le classement du fragment de document d_j (ranking) dans la liste finale des documents sélectionnés en réponse à la requête q .

4.3. RI temporelle

Le processus de RI pour la partie temporelle passe par les étapes suivantes. Tout d'abord, une requête est exprimée en texte libre. Cette requête est ensuite interprétée en utilisant la même chaîne de traitement que pour l'indexation. Soient Ens_{req} l'ensemble des entités marquées dans la requête et Ens_{doc} , l'ensemble des entités marquées dans le document. Nous avons : $Ens_{req} = \{ET_{req}\}$ et $Ens_{doc} = \{ET_{doc}\}$. Puis, nous calculons Ens_{res} qui est l'ensemble des entités de Ens_{doc} pour lesquelles l'intersection entre leur interprétation et celle d'une des entités de Ens_{req} n'est pas vide. Nous avons : $Ens_{res} = \{ET_{doc}\}$ avec $ET_{doc} \in Ens_{doc}$ et $\exists ET_{req} \in Ens_{req}$ tel que $interprete(ET_{doc}) \cap interprete(ET_{req}) \neq \emptyset$. Le résultat de la requête est l'ensemble des fragments de document auxquels les entités de Ens_{res} appartiennent. Un degré de pertinence est associé à chaque fragment résultat. Ce degré de pertinence est calculé en fonction du recouvrement entre l'intervalle de temps de l'ET du fragment de document et celui de l'ET de la requête.

Figure 4. Calcul de pertinence d'un document sélectionné

Pour chaque requête (voir figure 4), le système évalue le degré de similarité des fragments de document de la collection avec la requête. Tout d'abord, il s'agit de calculer le $Df_{precision}$ qui est égal au rapport de la durée commune au fragment de document évalué et à la requête ($I_{durée}$) avec la durée totale du fragment de document ($Df_{durée}$) : $Df_{precision} = \frac{I_{durée}}{Df_{durée}}$

Puis, le rapport avec durée de la requête ($Df_{significance}$) : $Df_{significance} = \frac{I_{durée}}{Q_{durée}}$

Et le rapport de distance : $Df_{distance} = \frac{d}{D}$

Ainsi, le degré de similarité Df_{score} est calculé comme suit :

$$Df_{score} = \frac{(Df_{precision} + Df_{significance})}{(2 + Df_{distance})} \quad [4]$$

Plus les centroïdes des durées de I et de Q sont proches, plus la pertinence du fragment de document est grande.

4.4. Combinaison de RI spatiale, temporelle et thématique (phénomène)

Nous avons déjà expérimenté un premier prototype combinant RI spatiale et thématique (Sallaberry *et al.*, 2007a). Les résultats suggèrent que d'un côté, l'approche spatiale est bien adaptée à la recherche de documents traitant de cette dimension mais l'est évidemment moins quand il s'agit de classer les documents pertinents dans les requêtes générales (espace + thème). L'approche classique, quant à elle, manque d'exhaustivité quand elle traite des requêtes spatiales, mais surclasse l'approche spatiale dans le cas de requêtes générales. Nous avons donc expérimenté une combinaison des deux approches pour améliorer le taux de précision des requêtes traitant de l'espace et du thème. Les détails des résultats obtenus sont donnés dans (Sallaberry *et al.*, 2007a).

Requêtes	P@5	P@10	P@15	Nombre de réponses
<i>Approche spatiale</i>				
Avg	0.15	0.18	0.18	1154
<i>Approche classique</i>				
Avg	0.48	0.39	0.36	331
<i>Combinaison des résultats de l'approche spatiale et classique</i>				
Avg	0.70	0.50	0.43	25.75

Tableau 4. Résultats de la combinaison de l'approche spatiale et classique

Les résultats confirment l'hypothèse de départ : combiner l'approche spatiale basée sur l'extraction et la comparaison des entités spatiales absolues et relatives avec l'approche thématique basée sur les statistiques permet d'améliorer les performances du système en classant plus de documents pertinents. Par exemple, aux cinq premiers documents (voir tableau 4), la précision atteint 70% quand les approches classique et spatiale sont combinées ; alors qu'elle n'est respectivement que de 48% et 15% quand les deux approches sont utilisées séparément. Cependant, on peut noter le nombre réduit de documents sélectionnés à cause de la méthode « prudente » de combinaison adoptée (intersection simple) : par exemple, dans le cas d'une de nos requêtes, l'approche combinée sélectionne seulement 4 unités documentaires alors que l'approche classique en retourne 233 et l'approche spatiale 724. L'amélioration de la précision au niveau des premiers documents restitués s'accompagne ainsi d'une diminution au niveau du rappel.

Nous envisageons désormais de combiner ces trois approches de RI d'une manière similaire à cette première expérimentation. L'idée est de subdiviser une requête en trois sous-requêtes (voir figure 5) ; la *sous-requête spatiale*, la *sous-requête temporelle* et la *sous-requête thématique*. Les sous-requêtes spatiales et temporelles contiennent les ES et les ET identifiées par les chaînes de traitement linguistique correspondantes. La sous-requête thématique contient les termes restants de la requête (phénomène). Comme schématisé dans la figure 5, prenons l'exemple d'une requête visant les « Instruments de musique dans les environs de Laruns au XIX^e siècle » : « les environs de Laruns » et « XIX^e siècle » représentent respectivement la sous-requête spatiale

et la sous-requête temporelle tandis que « *instrument* » et « *musique* » représentent la sous-requête thématique.

3

Figure 5. *Combinaison des approches*

Les trois sous-requêtes sont soumises au système d'appariement supportant l'approche appropriée. Le résultat final est construit en faisant une intersection des trois ensembles de documents sélectionnés (cf. figure 5). Le fait qu'une unité documentaire résultat corresponde à un paragraphe augmente la probabilité que les informations spatiales, temporelles et thématiques soient sémantiquement liées. Le classement final est basé sur celui obtenu à l'aide des algorithmes de calcul de similarité spatiale et temporelle.

5. Bilan et perspectives

Cette contribution se situe dans la lignée des approches d'extraction et de recherche d'information qui proposent des méthodes plus riches, atteignant le « sens » de manière plus fine, pouvant et devant être développées pour des espaces thématiques plus restreints et des tâches plus spécifiques. Une caractéristique majeure, peut-être même « la » caractéristique essentielle du « fait sémantique » est de rapporter une information à un espace de référence accepté ou posé à un moment déterminé par des usages.

Ce sont deux composantes d'un espace de référence, appelé « information géographique », que nous avons investies dans nos travaux. Cet espace convient bien, dans le cadre du projet PIV, dans la mesure où de nombreux contenus de documents (de la base documentaire de la médiathèque), y ancrent une part importante de leurs informations.

Les modèles sémantiques pour les composantes temporelles et spatiales de l'information géographique ont été élaborés sur un échantillon représentatif de ces documents. Cette manière de les concevoir fait leur force mais également leur limite. Issus du croisement entre l'étude du mode d'expression utilisé dans les documents, les objectifs et les impératifs d'un traitement automatique, ils sont tout à la fois légers (seuls les éléments pertinents sont modélisés), indépendants (les entités spatiales et les entités temporelles peuvent être traitées de manière totalement séparée) et compatibles entre eux. Mais leur généralité est limitée au cadre de corpus équivalents.

Comme indiqué dans la section 2.2 certains pré-requis et formalismes du langage TimeML, comme le besoin de connaître a priori l'événement (ce que nous avons nommé le phénomène/thème) ou alors l'absence de représentation associée aux relations temporelles, ne nous ont pas permis de l'utiliser efficacement dans notre problématique de RI.

Cette constatation nous a conduit à concevoir un langage de marquage, d'une part capable d'intégrer plus facilement les différents aspects de notre modèle, et d'autre part plus apte à répondre aux besoins de la phase de RI. A terme, il n'est pas à exclure qu'une telle démarche ne permette de proposer des extensions et/ou des évolutions du langage TimeML pour qu'il puisse intégrer ce type de besoin.

L'architecture choisie repose sur le principe des Services Web. Les modèles, les marquages sémantiques opérés sur les contenus ainsi que les index qui en découlent s'appuient sur la technologie XML. Les processus d'extraction et de recherche d'information sont entièrement implémentés. L'ensemble du processus d'extraction a fait l'objet d'une évaluation, il en résulte que plus de 80% des entités sont détectées et correctement décrites dans les modèles respectifs (des précisions, sur le mode opératoire de l'évaluation, pourront être trouvées dans (Sallaberry *et al.*, 2007b) et en particulier concernant les entités nommées spatiales). Dans le processus de recherche, seul l'aspect spatial, sur lequel ont porté nos premiers efforts, a été soumis à évaluation (la description du processus d'évaluation est décrit dans (Sallaberry *et al.*, 2007a)). Le protocole a montré que les résultats obtenus à partir de la combinaison des index spatiaux de PIV, avec les index thématiques, permettent un gain notable dans la pertinence des documents retournés. Une évaluation similaire est en cours concernant l'association des index temporels, détaillés dans cet article, avec les index thématiques. A terme une expérimentation en collaboration avec la Médiathèque est prévue afin de tester, dans un usage de type tourisme culturel, la qualité des réponses retournées lorsqu'elles sont issues de l'intersection des trois index.

Néanmoins, le problème de fusion des trois ensembles de résultats obtenus dans la phase de RI pour optimiser aussi bien la précision que le rappel, reste un problème ouvert. Ceci nécessite l'étude d'opérateurs plus complexes (union, avec intégration pondérée de la pertinence spatiale, temporelle et statistique, par exemple) que la simple intersection utilisée ici. Les prochains travaux devront mener à des propositions concernant le calcul du degré global de pertinence d'une unité documentaire retournée ; les algorithmes d'intégration des taux de pertinence spatial, temporel et thématique devront tenir compte du contexte et des modes de calcul spécifiques de chacun de ces

trois ratios. Des expérimentations d'algorithmes de combinaisons spatiale et thématique utilisant le produit, la similarité maximum, des fonctions de combinaison linéaire ont, par exemple, été décrites dans (Martins *et al.*, 2005).

6. Bibliographie

- Allen J., « Towards a General Theory of Action and Time », *Artificial Intelligence*, 1984.
- Aussenac-Gilles N., Jacques M.-P., « Variabilité des performances des outils de TAL et genre textuel. Cas des patrons lexico-syntaxiques », *Traitement Automatique des Langues, Numéro spécial Non Thématique*, 2006.
- Baeza-Yates R. A., Ribeiro-Neto B. A., *Modern Information Retrieval*, ACM Press / Addison-Wesley, 1999.
- Bilhaut F., « The Linguastream Platform », *Proceedings of the 19th Spanish Society for Natural Language Processing Conference (SEPLN)* p. 339-340, 2003.
- Boughanem M., Chrisment C., Tmar M., « Mercure and MercureFiltre Applied for Web and Filtering Tasks at TREC-10 », *TREC*, 2001.
- Bourigault D., Jacquemin C., L'Homme M.-C., (eds), *Recent Advances in Computational Terminology*, John Benjamins, Amsterdam, 2001.
- Chen Y.-Y., Suel T., Markowetz A., « Efficient query processing in geographic web search engines », *ACM SIGMOD international conference on Management of data*, p. 277-288, 2006.
- Enjalbert P., Gaio M., « Actualité d'une approche sémantique du document numérique », *Actes du 7ième Colloque International sur le Document Electronique (CIDE.7)*, 2004.
- Ferro L., Gerber L., Mani I., Sundheim B., Wilson G., TIDES - 2003 Standard for the Annotation of Temporal Expressions, Technical report, 2003.
- Gaio M., Sallaberry C., Etcheverry P., Marquesuzaà C., Lesbegueries J., « A Global Process to Access Documents' Contents from a Geographical Point of View », *Journal Of Visual Languages And Computing*, à paraître en décembre 2007. Elsevier.
- Gosselin L., *Sémantique de la temporalité en français. Un modèle calculatoire et cognitif du temps et de l'aspect*, Duculot, 1996.
- Lemmens R., « Exploitation of ontology mappings for the discovery of geo - web services », *Geographic information science : proceedings of the 4th international conference GIS-science 2006*, 2006.
- Lesbegueries J., Gaio M., Loustau P., Sallaberry C., « Geographical information access for non structured data », *ACM SAC - Advances in Spatial and Image based Information Systems track*, vol. 1, ACM Press, Dijon (FR), p. 83-89, Avril, 2006.
- Mani I., « Recent Developments in Temporal Information Extraction », *RANLP'03*, 2004.
- Mani I., Wilson G., « Temporal granularity and temporal tagging of text », *AAAI-200 Workshop on Spatial and Temporal Granularity*, 2000.
- Martins B., M. Silva M.-J., Andrade L., « Indexing and ranking in Geo-IR systems », *2nd Int. Workshop on Geo-IR (GIR)*, 2005.
- MUC-7, « Proceedings of the Seventh Message Understanding Conference », 1998.

- Muller P., Tannier X., « Annotating and measuring temporal relations in texts », *Coling 2004*, Genève, Association for Computational Linguistics, p. 50-56, 2004.
- Porter M., « An algorithm for suffix stripping », *Program*, 1980.
- Robertson S. E., Walker S., Hancock-Beaulieu M., « Automatic Ad Hoc, Filtering, VLC and Interactive », *TREC*, 1998.
- Sagara T., Kitsuregawa M., « Yellow Page driven Methods of Collecting and Scoring Spatial Web Documents », *SIGIR Workshop on Geographical Information Retrieval*, 2004. <http://www.geo.unizh.ch/rsp/gir/>.
- Sallaberry C., Baziz M., Lesbegueries J., Gaio M., « Towards an IE and IR system dealing with spatial information in digital libraries : Evaluation Case Study », *9th International Conference on Enterprise Information Systems - Human-Computer Interaction Area / Geographical Information Systems, ICEIS*, 2007a. <http://snowball.tartarus.org/texts/introduction.html>.Ricardo.
- Sallaberry C., Gaio M., Lesbegueries J., Loustau P., « A Semantic Approach for Geospatial Information Extraction from Unstructured Documents », *Geospatial Web*, 2007b. ISBN 1-84628-826-6.
- Sanderson M., Kohler J., « Analyzing geographic queries », *Proceedings of the Workshop on Geographic Information Retrieval, SIGIR*, 2004.
- Sauri R., Littman J., Knippen B., Gaizauskas R., Setzer A., Pustejovsky J., TimeML Annotation Guidelines - Version 1.2.1, Technical report, 2006.
- Schilder F., Habel C., « From temporal expressions to temporal information : semantic tagging of news messages », *Proceedings of the ACL 2001 Workshop on Temporal and Spatial Information Processing*, Toulouse, p. 65-72, 2001.
- Schmid H., « Probabilistic part-of-speech tagging using decision trees », *International Conference on New Methods in Language Processing*, 1994.
- Setzer A., Gaizauskas R., « On the Importance of Annotating Event-Event Temporal Relations in Text », *LREC 2002, Workshop on Annotation Standards for Temporal Information in Natural Language*, 2002.
- Vaid S., Jones C. B., Joho H., Sanderson M., « Spatio-textual indexing for geographical search on the web », *9th Int. Symp. on Spatial and Temporal Databases (SSTD)*, 2005.
- Vazov N., « A System for Extraction of Temporal Expressions from French Texts Based on Syntactic and Semantic Constraints », *Workshop on Temporal and Spatial Information Processing*, 2001.
- Widlocher A., Bilhaut F., « La plateforme linguastream : un outil d'exploration linguistique sur corpus », *12e conférence Traitement Automatique du Langage Naturel*, 2005.
- Wilson G., Mani I., Sundheim B., Ferro L., « A Multilingual Approach to Annotating and Extracting Temporal Information », *Workshop on Temporal and Spatial Information Processing*, 2001.