Réconcilier les événements dans le web de données

Houda Khrouf, Raphaël Troncy

Réconcilier les événements dans le web de données

Houda Khrouf et Raphaël Troncy

EUROCOM, Sophia Antipolis, France
<houda.khrouf@eurecom.fr>
<raphael.troncy@eurecom.fr>

Résumé : De nombreux sites web fournissent des informations à propos d'événements passés ou à venir, et certains d'entre eux affichent même des photos ou des vidéos capturés pendant ces événements. L'information disponible est, cependant, souvent incomplète, erronée et enfermée dans une multitude de sites web. Notre objectif est de fournir une application web permettant de revivre ou de découvrir des événements à partir de médias. Dans cet article, nous avons tout d'abord cherché à évaluer la quantité et la qualité des jeux de données disponibles sur le web sémantique contenant des informations liées à des événements. Nous montrons alors comment nous avons réconcilié d'importants volumes de données. Nous décrivons les problèmes rencontrés et présentons quelques défis pour les outils d'interconnexion des données.

Mots-clés : LODE, EventMedia, réconciliation de données, alignement, web sémantique

1. Introduction

Dans des travaux précédents, nous avons effectué plusieurs études centrées utilisateur pour mieux comprendre comment les utilisateurs découvraient et participaient à des événements ou partageaient cette expérience, et quels outils ils utilisaient pour cela (Fialho *et al.*, 2010). Le résultat de ces études soutient l'idée de développer une application web qui agrégerait des informations disponibles dans des annuaires d'événements avec des témoignages média capturés par les utilisateurs tout en offrant des fonctionnalités sociales (Troncy *et al.*, 2010a). Notre postulat est que les technologies du web sémantique sont adaptées pour effectuer l'intégration à large échelle de toutes ces données.

Le web de données [1] a en effet comme double objectif de *i)* publier des descriptions représentées en RDF dont les URIs identifient des documents web, des objets du monde réel et des relations les reliant et *ii)* d'interconnecter ces jeux de données. A l'issue de processus sociaux, certains vocabulaires sont devenus très populaires facilitant ainsi l'interconnexion des données (Vatant & Rozat, 2011). Ainsi, on utilisera plutôt le vocabulaire `Dublin Core` [2] pour attacher un titre ou une description à une ressource, `FOAF` [3] pour décrire une personne ou un groupe et `WGS84` [4] pour représenter les lieux géographiques. Pourtant, nous observons qu'aucun vocabulaire n'a encore véritablement émergé pour représenter la notion d'événement.

Dans cet article, nous avons tout d'abord cherché à évaluer la quantité et la qualité des jeux de données disponibles sur le web sémantique contenant des informations liées à des événements. Nous avons alors construit le jeu de données EventMedia composé d'une part de descriptions sémantiques d'événements et d'autre part de photos et vidéos illustrant ceux-ci. Nous avons mis en correspondance manuellement les modèles et référentiels sous-jacents (section 2.). Nous décrivons ensuite comment nous avons interconnecté plusieurs noeuds centraux du web de données avec EventMedia (section 3.). Nous discutons des problèmes rencontrés pour réconcilier et nettoyer ces données qui posent de vrais défis pour les outils d'interconnexion des données du web sémantique (section 4.) avant de conclure et d'ouvrir quelques perspectives à ces travaux (section 5.).

2. Quelle bulle du web de données contient des événements ?

Le terme "événement" est polysémique : il fait tout à la fois référence à des phénomènes passés (décrits dans des articles de presse ou expliqués par des historiens) et à des phénomènes planifiés dans le futur (notés dans un calendrier ou une programmation). Dans des travaux précédents, nous avons analysé les différentes ontologies permettant de représenter la notion d'événement. Nous avons alors proposé l'ontologie LODE qui fourni un modèle simple pour représenter les différentes propriétés composant un événement ainsi qu'un ensemble de correspondances entre de nombreux modèles pour le

1. `http://linkeddata.org/`
2. `http://purl.org/dc/elements/1.1`
3. `http://xmlns.com/foaf/0.1`
4. `http://www.w3.org/2003/01/geo/wgs84_pos`

représenter (Troncy *et al.*, 2010c).

Dans cette section, nous présentons d'abord l'ontologie LODE à l'aide d'un exemple (section 2.1.). Nous décrivons ensuite la fabrication du jeu de données EventMedia qui a fait son apparition dans le nuage de données du web sémantique (section 2.2.). Nous identifions enfin quel jeu de données peut être interconnecté avec EventMedia et comment l'alignement des modèles et référentiels utilisés a été obtenu (section 2.3.).

2.1. LODE par l'exemple

LODE [5] est une ontologie minimale permettant la description interopérable des aspects "factuels" d'un événement, ce qui peut se caractériser en terme des "quatre Ws" (*what*, *when*, *where*, *who*) : qu'est-ce qui s'est passé, où et quand cela s'est-il produit, qui était impliqué. Ces relations factuelles décrivant un événement ont comme objectif de représenter une réalité consensuelle et ne doivent donc pas être associées à une perspective ou une interprétation particulière.

La figure 1 illustre comment l'événement identifié par `350591` sur last.fm serait décrit avec l'ontologie LODE. Plus précisément, elle montre qu'un événement de type `Concert` a été donné le `13 juillet 2007 à 20h30` au théâtre le `Nouveau Casino` à Paris avec comme vedette la chanteuse irlandaise `Róisín Murphy` connue pour sa musique électronique.

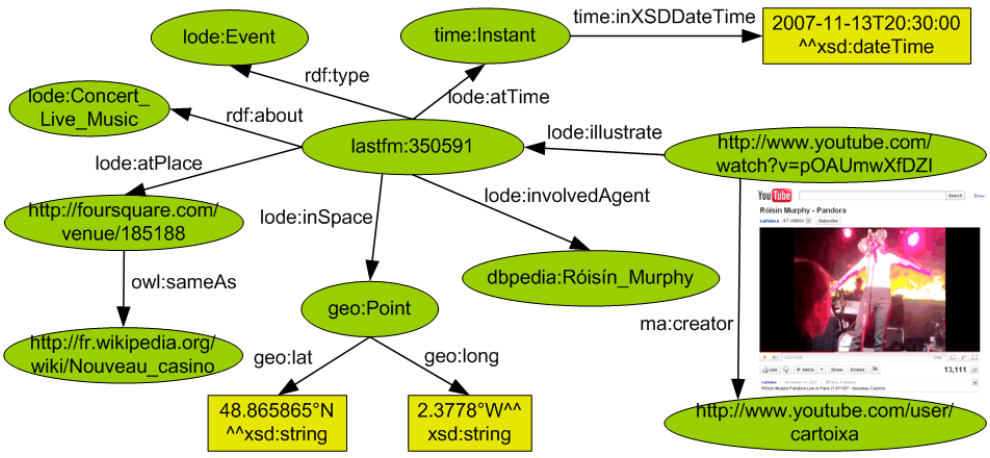

FIGURE 1: *Róisín Murphy au Nouveau Casino à Paris* décrit avec LODE

5. `http://linkedevents.org/ontology/`

2.2. EventMedia = Last.fm + Eventful + Upcoming + Flickr

EventMedia est une nouvelle bulle [6] du nuage de données apparue dans sa représentation imagée publiée en septembre 2010 (Cyganiak & Jentzsch, 2010). EventMedia est composée de descriptions d'événements publiés sur Last.fm, Eventful et Upcoming qui ont au moins une photo publiée sur Flickr explicitement associée à ces événements.

Nous utilisons les APIs de ces trois annuaires d'événements pour convertir les descriptions selon l'ontologie LODE. Nous créons nos propres URIs dans notre espace de noms pour représenter les événements (`http://data.linkedevents.org/event/`), les agents (`http://data.linkedevents.org/agent/`) et les lieux (`http://data.linkedevents.org/location/`). Un graphe représentant un événement est ainsi composé du type de l'événement, d'une description textuelle, des personnes impliquées, d'une date (un instant ou un intervalle représenté avec l'ontologie OWL Time (Hobbs & Pan, 2006)), d'un lieu représenté à la fois en terme de coordonnées géographiques et d'une étiquette. Un graphe représentant un agent ou un lieu contient une étiquette et une description textuelle (e.g. la biographie d'un artiste), le lieu ayant en plus une adresse structurée.

Une relation explicite entre un événement publié dans un annuaire et une photo hébergée sur Flickr peut être retrouvée à l'aide de tags sémantiques spéciaux tels que `lastfm:event=XXX` ou `upcoming:event=XXX`. Dans un travail précédent, nous avons collecté l'intersection des sites Last.fm, Eventful et Upcoming avec Flickr pour obtenir un jeu de données composé de plus de 140 000 descriptions d'événements associés à plus de 1,7 million de photos (Tableau 1).

	Event	**Agent**	**Location**	**Photos**	**User**
`Last.fm`	37,647	50,151	16,471	1,393,039	18,542
`Upcoming`	13,114	-	7,330	347,959	4,518
`Eventful`	37,647	6,543	14,576	52	12

TABLE 1: Volume de descriptions pour les classes event/agent/location et photo/user dans le jeu de données EventMedia (Troncy *et al.*, 2010b)

6. Voir aussi la description sur CKAN à `http://ckan.net/package/event-media`

2.3. Aligner les ontologies et les référentiels

Outre EventMedia, d'autres bulles du nuage de données contiennent des informations liées à des événements, des personnes ou des lieux. Nous avons tout d'abord effectué un certain nombre de requêtes génériques sur le cache du web de données fournit par Virtuoso à (`http://lod.openlinksw.com/`) puis des requêtes dédiées sur des points SPARQL précis liés à certains jeux de données. Ainsi, la classe `Event` est définie dans les vocabulaires Yago, DBpedia, Uberblic, Cyc, Umbel, etc.

Au final, nous avons repéré les bulles suivantes susceptibles de contenir des données recouvrant EventMedia :

- pour la classe `Agent` : Last.fm, Eventful, MusicBrainz, DBpedia, Freebase, Uberblic, New York Times
- pour la classe `Location` : Last.fm, Eventful, Upcoming, DBpedia, Freebase, Foursquare, Geonames
- pour la classe `Event` : Last.fm, Eventful, Upcoming, DBpedia, Freebase, Uberblic

Nous décrivons dans la suite comment nous avons manuellement aligné les ontologies et les référentiels décrivant ces jeux de données en préalable à l'interconnexion des données.

2.3.1. Les modèles d'événements

L'ontologie LODE est utilisée pour décrire les événements publiés dans EventMedia. L'ontologie Event (Raimond *et al.*, 2007) constitue elle la base de MusicBrainz. Les jeux de données encyclopédiques issues de Wikipedia tels que DBpedia, Freebase ou Uberblic définissent leur propre classe d'événement en la spécialisant selon leur type. Nous verrons dans la suite comment nous avons aligné ces catégories (ou référentiels) d'événements.

LODE a été conçu à la base comme un modèle interopérable pour décrire des événements. Ainsi, il fournit un ensemble d'axiomes logiques entre de nombreuses classes et propriétés définies dans d'autres modèles tels que les ontologies Event, CIDOC-CRM, DOLCE, SEM (van Hage *et al.*, 2009) pour n'en citer que quelques unes. Le tableau 2 montre quelques unes des propriétés du modèle LODE avec leurs correspondances.

ABC	CIDOC	DUL	EO	LODE
atTime	P4.has_time-span	isObservableAt	time	atTime
	P7.took_place_at		place	inSpace
inPlace		hasLocation		atPlace
involves	P12.occurred_in_-the_presence_of	hasParticipant	factor	involved
hasPresence	P11.had_-participant	involvesAgent	agent	involvedAgent

TABLE 2: Exemple d'alignements entre propriétés de plusieurs ontologies événements

2.3.2. Les catégories d'événements

Les événements sont généralement catégorisés en taxonomies qui fournissent sur de nombreux sites un moyen pratique de parcourir les événements publiés par type. Nous avons manuellement analysé les taxonomies proposées par différents sites tels que : Facebook, Eventful, Upcoming, Zevents, LinkedIn, EventBrite, TicketMaster ainsi que les bulles encyclopédiques du nuage de données. Nous avons alors appliqué la technique du tri par cartes [7] pour construire un thésaurus de catégories d'événements contenant des renvois à ces sources. Le thésaurus est représenté en SKOS et les termes sont définis dans notre espace de noms (`http://data.linkedevents.org/category/`). Les catégories de haut niveau sont ainsi : Sports, Music, Food, Arts, Movies, Family, Social Gathering, Community, Professional et Military Conflicts. Nous fournissons également un alignement avec d'autres classifications telles que les News Codes définis par l'IPTC [8] pour les différents sports, ou les genres définis par Last.fm pour la musique.

3. Réconciliation des données

Après avoir identifié les bulles du nuage de données susceptibles de contenir des informations liées à des événements, nous avons cherché à interconnecter à large échelle ces jeux de données. Pour le jeu de données EventMedia, nous avons chargé dans un serveur Virtuoso local les données dans trois graphes distincts selon leur provenance (Last.fm, Eventful ou Upcoming).

7. `http://fr.wikipedia.org/wiki/Tri_par_cartes`

8. `http://cv.iptc.org`

Pour les autres jeux de données, nous avons utilisés les points SPARQL offerts par ceux-ci.

Nous commençons par décrire le framework Silk et sa configuration (section 3.1.), puis la méthodologie d'alignement utilisé (section 3.2.). Nous présentons ensuite les résultats obtenus pour l'interconnexion des agents (section 3.3.), des lieux (section 3.4.) et des événements selon leur genre (section 3.5.).

3.1. Le framework d'alignement Silk

Le framework Silk (Volz *et al.*, 2009a,b) fournit un moteur d'interconnexion de données en se basant sur le langage de spécification de liens (Silk-LSL) qui définit les conditions et restrictions à appliquer pour l'appariement. Silk permet d'accéder à des sous ensembles de données hébergées sur des points SPARQL. Il compare alors les descriptions de jeux de données source (`<SourceDataset>`) et cible (`TargetDataset`) à l'aide de mesures de similarités. Le langage de spécification permet de définir quelles métriques utiliser et comment les combiner à l'aide de fonctions algébriques (e.g. minimum, maximum, moyenne). De plus, Silk fournit un langage de sélection dans des chemins d'un graphe RDF pour suivre les ressources. Finalement, des fonctions permettant de manipuler les chaîne de caractères telles que normaliser la casse, changer l'encodage, remplacer les espaces ou certains caractères, complètent le framework.

De nombreuses mesures de similarités pour aligner des ontologies ont été proposées (Euzenat & Shvaiko, 2007). Le langage Silk-LSL permet d'utiliser des métriques syntaxiques (e.g. égalité, Jaro, Levenstein), lexicales (e.g. WordNet), temporelles (e.g. date) ou géographiques (e.g. wgs84) définies comme suit :

- equality : retourne 1 si les chaînes de caractères sont identiques et 0 sinon ;
- q-grams : compte le nombre de q-grams (un ensemble de sous-chaînes de caractères de longueur q) en commun entre deux chaînes (des caractères supplémentaires peuvent être ajoutés en début ou en fin de chaînes quand celles-ci sont trop courtes) ;
- Jaro : compte le nombre de caractères communs et leurs transposition entre deux chaînes de caractères ;
- Jaro-Winkler : modifie la mesure de Jaro en donnant un poids plus élevé aux préfixes communs ;

– wgs84 : calcule la distance géographique entre deux points définis à l'aide d'une latitude et d'une longitude.

Une fois la configuration spécifiée, Silk calcule un score d'appariement pour chaque ressource. Ce score est ensuite seuillé pour filtrer les interconnexions jugées valides qui sont alors exportées sous forme de liens `sameAs`. La figure 2 illustre un bloc de conditions pour l'appariement en précisant la fonction d'agrégation (maximum) et la mesure de similarité (jaro) à utiliser après avoir transformé la casse des noms de ressources.

```
<LinkCondition>
 <Aggregate type="max">
  <Compare metric="jaro">
   <TransformInput function="lowerCase">
    <Input path="?a/rdfs:label"/>
   </TransformInput>
   <TransformInput function="lowerCase">
    <Input path="?b/rdfs:label"/>
   </TransformInput>
  </Compare>
  <Compare metric="wgs84">
   <TransformInput function="concat">
    <Input path="?a/lode:atPlace/lode:inSpace/wgs84:lat"/>
    <Input path="?a/lode:atPlace/lode:inSpace/wgs84:long"/>
    <Param name="glue" value=" "/>
   </TransformInput>
   <TransformInput function="concat">
    <Input path="?b/lode:atPlace/lode:inSpace/wgs84:lat"/>
    <Input path="?b/lode:atPlace/lode:inSpace/wgs84:long"/>
    <Param name="glue" value=" "/>
   </TransformInput>
   <Param name="unit" value="km"/>
   <Param name="threshold" value="10"/>
  </Compare>
 </Aggregate>
</LinkCondition>
<Filter threshold="0.8"/>
```

FIGURE 2: Exemple d'un bloc de conditions pour l'appariement de ressources dans Silk

3.2. Méthodologie d'alignement

Compte tenu de la nature des données représentées dans EventMedia (personnes, lieux et événements), deux types de ressources peuvent être utilisées pour l'appariement : les noms des ressources et les coordonnées géographiques dans le cas des lieux. Certaines ressources ont en outre une description (e.g. la biographie d'un artiste). Cependant, celle ci contient parfois des caractères non alpha-numériques, s'avère souvent longue et est généralement très variable. Les mesures de similarité entre chaînes de caractères ont tendance à bruiter fortement les résultats d'alignement dans ce cas. Une mesure de distance entre textes normalisés serait sans doute plus appropriée mais Silk n'est pas encore équipé d'une telle métrique.

3.2.1. Alignement par les étiquettes

Les noms des ressources publiées à la fois dans EventMedia et dans les jeux de données identifiés précédemment sont généralement disponibles en anglais. Ainsi, nous pouvons utiliser des mesures de similarité purement syntaxiques sans avoir à manipuler des ressources lexicales telles que WordNet. Intuitivement, la mesure d'égalité semble suffisante pour aligner des personnes, lieux ou événements. Cependant, l'emploi de nombreux caractères de ponctuation (apostrophe, deux points, parenthèse) ou accentués viennent bruiter les résultats. Plutôt que d'imaginer quels pourraient être tous les caractères spéciaux utilisés afin de les normaliser, nous avons opté pour une agrégation entre différentes mesures syntaxiques à l'aide de l'opérateur maximum. La mesure q-grams ne donnant pas de bons résultats sur les chaînes de caractères courtes et la mesure Jaro-Winkler donnant un biais trop important en cas de préfixes communs (Euzenat & Shvaiko, 2007), nous avons appliqué la mesure simple de Jaro (équation 1).

$$sim(l_1, l_2) = MAX(equality(l_1, l_2), jaro(l_1, l_2)) \quad (1)$$

Nous avons effectué différents tests pour définir la valeur de seuil validant les appariements pour finalement choisir 0,98. Cette valeur de seuil, plutôt haute, est conservatrice : elle permet d'apparier deux chaînes qui diffèrent d'un caractère au maximum sans introduire trop de bruit. Ainsi, les événements `Shipley Open-mic Feature` définis dans Upcoming et `Shipley Open mic Feature` définis dans Eventful sont alignés. En revanche, nous avons aussi constaté certains cas où l'appariement ne devrait pas se faire. Ainsi,

la bataille de Monterey définit dans Uberblic à (`http://uberblic.org/resource/78cab524-a012-45e2-8f25-55b9dc09fc23#thing`) est différente de la bataille de Monterrey définit dans DBpedia (`http://dbpedia.org/page/Battle_of_Monterrey`) bien qu'ayant toutes les deux eu lieu en 1846 pendant la guerre opposant le Mexique et les États Unis puisque elles ont été conduites respectivement par les généraux John D. Sloat et Zachary Taylor. Nous avons estimé que la probabilité de rencontrer deux événements avec autant de propriétés communes était faible par rapport au gain d'avoir un caractère flottant dans la comparaison de chaînes.

3.2.2. Alignement par les coordonnées géographiques

Un alignement purement basé sur la comparaison des noms donne de très mauvais résultats pour l'appariement des lieux. Par exemple, 8 lieux différents ont tous le même nom `Starbucks` dans EventMedia. La définition structurée de l'adresse quand elle existe ou les coordonnées géographiques peuvent aussi être utilisés. De manière empirique, nous avons établi qu'une combinaison pondérée entre une distance sur les chaînes de caractères et une distance linéaire entre deux points géographiques donnaient les meilleurs résultats (équation 2). Nous discutons de la valeur 6 donnée au poids de la mesure de similarité géographique dans la section 4.

$$sim(r_1, r_2) = MAX(equality(l_1, l_2), jaro(l_1, l_2)) + 6 * wgs84(p_1, p_2) \quad (2)$$

ou r_1 (resp. r_2) a comme nom l_1 (resp. l_2) et comme point p_1 (resp. p_2)

Nous détaillons dans la suite les résultats obtenus pour interconnecter les personnes, les lieux et les événements d'EventMedia avec le nuage de données selon cette méthodologie.

3.3. Alignement des agents

Les jeux de données pertinents pour aligner les personnes sont : MusicBrainz qui contient une grande base de données d'artistes et d'albums, DBpedia, Freebase et Uberblic tous trois générés à partir de Wikipedia et New York Times qui contient un référentiel de personnes pour lesquels des articles de presse ont été écrit. Dans EventMedia, les agents sont de type `foaf:Agent` et ont un nom identifié par la propriété `rdfs:label`. Le résultat des appariements entre tous ces jeux de données et ceux de Last.fm et Eventful disponibles dans EventMedia sont résumés dans le tableau 3. Les nombres entre parenthèses indiquent le nombre total d'instances susceptibles d'être alignées.

	Eventful (6543)	**Last.fm** (50151)	**MusicBrainz** (459023)	**DBpedia** (107112)	**Uberblic** (236691)	**NYTimes** (4794)
Eventful	-	2865 (44%)	3616 (55%)	1985 (30%)	1567 (24%)	7 (0.1%)
Last.fm	2865 (6%)	-	26619 (53%)	9442 (19%)	12905 (26%)	14 (0.03%)

TABLE 3: Alignements des agents entre plusieurs jeux du nuage de données obtenus avec l'équation 1

Cette première expérimentation nous confirme l'utilité de la métrique Jaro. Celle-ci apporte 18 appariements supplémentaires entre les jeux de données Eventful et Last.fm parmi lesquels `Antipop Consortium` et `anti-pop consortium`), `Donavon Frankenreiter` et `Donovan Frankenreiter` ou encore `Hawthorne Heights` et `Hawthore Heights`. Ainsi, la réconciliation des données permet de mettre en évidence les erreurs typographiques entrées par les utilisateurs. Au total, nous avons réussi à aligner 4014 (soit 61%) des personnes définies par Eventful et 29138 (soit 58%) des personnes définies par Last.fm avec au moins un autre jeu de données.

3.4. Alignement des lieux

Les jeux de données pertinents pour aligner les lieux sont : Foursquare qui fournit une gigantesque base de données de points d'intérêts et dont les descriptions ont été partiellement converties en RDF dans Uberblic, DBpedia et Freebase tous deux générés à partir de Wikipedia, et Geonames. Dans EventMedia, les trois sous-ensembles constitués à partir de Eventful, Last.fm et Upcoming contiennent des descriptions géo-localisées de lieux. Le résultat des appariements entre tous ces jeux de données et ceux de EventMedia sont résumés dans le tableau 4. Les nombres entre parenthèses indiquent le nombre total d'instances susceptibles d'être alignées.

Cette deuxième expérimentation nous confirme la nécessité d'utiliser la métrique wgs84. Par exemple, les lieux `The Stone Bar` et `The Stone` ne passent pas le seuil de 0,98 avec la seule métrique de Jaro. La distance linéaire calculée avec leurs coordonnées – (34.1019 ;-118.304) et (34.1017503 ;-118.3042771) – permet en revanche de les apparier. Nous observons que le jeu de données Foursquare est extrêmement bruité dans la mesure où il contient lui même de nombreux doublons puisque les points d'intérêts peuvent être ajoutés à tout moment par n'importe quel utilisateur et associés à des coordonnées géographiques variables et parfois largement erronées selon la fia-

	Eventful (13516)	**Last.fm** (15857)	**Upcoming** (5173)	**DBpedia** (496728)	**Foursquare** (641770)	**Geonames** (1090357)
Eventful	-	998 (7%)	366 (3%)	90 (0,7%)	1296 (10%)	320 (2%)
Last.fm	998 (6%)	-	626 (4%)	141 (0.9%)	911 (6%)	345 (2%)
Upcoming	366 (7%)	626 (12%)	-	74 (1,4%)	1300 (25%)	232 (4%)

TABLE 4: Alignements des lieux entre plusieurs jeux du nuage de données obtenus avec l'équation 2

bilité des GPS embarqués sur les appareils mobiles. Enfin, les résultats d'alignement avec Geonames ne sont que partiels puisque seul un million d'entités (16%) ont pu être considérées. Une manière de passer l'échelle est d'utiliser Silk sur MapReduce [9] basé sur Hadoop ce qui requiert d'utiliser (ou de louer) un cluster de machines. Au total, nous avons réussi à aligner 2292 (soit 17%) des lieux définis par Eventful, 2384 (soit 15%) des lieux définis par Last.fm et 1870 (soit 36%) des lieux définis par Upcoming avec au moins un autre jeu de données.

3.5. Alignement des événements

Les jeux de données pertinents pour aligner les événements sont d'une part Eventful, Last.fm et Upcoming disponibles dans EventMedia et d'autre part DBpedia, Freebase et Uberblic tous trois générés à partir de Wikipedia. Un événement est généralement complètement décrit à l'aide d'un titre, d'un lieu et d'un intervalle de temps. Nous avons donc testé différentes combinaisons basées sur ces propriétés pour aligner des descriptions d'événements. Dans la suite, nous présentons les résultats des appariements entre tous ces jeux de données selon le type de l'événement.

3.5.1. Aligner les événements musicaux

Bien qu'EventMedia contienne des descriptions pour des événements de type très différents (concerts, festivals, photographie, conférence, technologie, exposition...) seuls les événements de types musicaux semblent avoir des correspondances avec les jeux de données issus de Wikipedia. Les résultats des appariements entre DBpedia et EventMedia sont résumés dans le tableau 5.

9. `http://www4.wiwiss.fu-berlin.de/bizer/silk/mapreduce/`

Les nombres entre parenthèses indiquent le nombre total d'instances susceptibles d'être alignées.

	Eventful (37647)	**Last.fm** (57258)	**Upcoming** (13114)	**DBpedia** *Music Festival* (662)	**Uberblic** *Performer* (228238)
Eventful	-	76 (0,2%)	34 (0,1%)	28 (0,1%)	15 (0,04%)
Last.fm	76 (0,1%)	-	586 (1%)	389 (0,7%)	1148 (2%)
Upcoming	34 (0,3%)	586 (4%)	-	31 (0,2%)	15 (0,1%)

TABLE 5: Alignements des événements musicaux entre plusieurs jeux du nuage de données

Nous avons constaté que l'appariement sur les titres seuls étaient peu fiable. Ainsi, l'ensemble des événements décrits dans Last.fm auraient au moins une correspondance dans Upcoming si l'on ne se fie qu'au titre des événements. Le titre et le lieu ne sont parfois pas suffisants puisque certains événements sont récurrents. Au final, les chiffres indiqués dans le tableau 5 correspondent à une mesure de similarité qui prend en compte le titre de l'événement, son lieu et sa date. La mesure de comparaison des dates a été personnalisé puisque la mesure de Silk est par défaut trop rigide, ne tenant pas compte des fuseaux horaires ou de l'inclusion temporelle. Ainsi, l'exposition `A Season in Hell` a eu lieu du 7 novembre au 22 novembre 2008 d'après Upcoming (`http://upcoming.yahoo.com/event/1326644`) alors qu'elle s'est déroulée uniquement le 8 novembre d'après Eventful (`http://eventful.com/events/E0-001-017164274-1@2008110812`). Au total, nous avons réussi à aligner 139 (soit 0,4%) des événements définis par Eventful, 2163 (soit 3,8%) des événements définis par Last.fm et 626 (soit 4,8%) des événements définis par Upcoming avec au moins un autre jeu de données.

3.5.2. Aligner d'autres types d'événements

EventMedia étant construit à partir de Last.fm, Eventful et Upcoming, il ne contient aucun événement de type sportif, conflit militaire ou mission spatiale. En revanche, les jeux de données DBpedia, Freebase et Uberblic étant tous trois issus de Wikipedia, on retrouve une interconnexion forte entre ces sous-ensembles (tableau 6).

Compte tenu de leur origine commune, il n'est pas surprenant de retrouver une interconnexion forte entre ces jeux de données. Une exception ap-

Type	**DBpedia**	**Uberblic**	**Alignement**
Military conflict	8 750	8 899	7 151 (81%)
Space Mission	396	362	346 (95%)
Sport Events	4 046	3 056	942 (30%)

TABLE 6: Alignements des événements de type conflit militaire, mission spatiale et sportif entre plusieurs jeux du nuage de données

paraît toutefois pour les événements sportifs où l'on retrouve seulement 30% d'appariement. L'explication est liée à la manière dont les ressources sont nommées. Dans DBpedia, le nom de l'événement contient généralement une indication de la date (voire du lieu) où il a eu lieu alors que Uberblic génère des noms plus canoniques. Les mesures de similarités syntaxiques obtiennent donc un mauvais score. Il faut alors soit baisser la valeur de seuil permettant de valider les appariements, soit donner une importance très faible à cette distance par rapport au lieu et à la date de l'événement avec le risque d'introduire beaucoup de bruit dans les alignements.

4. Résultats et discussion

Les alignements sont-ils corrects et complets ? La quasi totalité des alignements trouvés ont été validé manuellement. Un petit nombre d'erreurs a été constaté mais globalement la précision des alignements est supérieure à 90%. Ceci s'explique par l'approche très conservatrice que l'on a suivi avec un seuil très haut pour filtrer les appariements. Il est en revanche plus compliqué d'évaluer la complétude des alignements obtenus pour les événements. En fait, les résultats diffèrent selon les ressources alignées. Pour les personnes, peu d'appariements sont manquants. Pour les lieux, un gros travail de nettoyage doit être effectué, en particulier pour manipuler des jeux de données telle que Foursquare qui est extrêmement riche mais contient beaucoup de doublons. Une mesure de similarité prenant exclusivement en compte la structure d'une adresse est sans doute à étudier. Enfin, l'appariement des événements donne des résultats contrastés. L'erreur la plus fréquente apparaît dans le cas d'événements récurrents où un événement particulier est aligné avec une page générique de disambiguisation dans DBpedia.

Quel rayon géographique choisir pour l'alignement des lieux ? Les capteurs GPS des appareils mobiles ont une certaine marge d'erreur. Dans

des environnements urbains à forte densité, il n'est pas rare de trouver deux événements qui ont lieu en même temps dans un périmètre très proche. Nous avons effectué plusieurs expériences pour calculer quel devait être le poids de la distance wgs84 par rapport à la mesure de similarité syntaxique sur les noms des lieux dans l'équation 2. Par exemple, le lieu `American Airlines Center` est présent deux fois dans Upcoming (`http://upcoming.yahoo.com/venue/89278` et `http://upcoming.yahoo.com/venue/35660`) au point (32.7922,-96.8069) ou (32.7863,-96.7974) soit distants de 1,104 km ! Pour apparier ces deux lieux, il faudrait que la mesure wgs84 compte 60 fois plus que la mesure de similarité syntaxique si l'on souhaite garder comme seuil la valeur 0,98. Le paramètre de poids 6 correspond à une marge d'erreur de 10 km pour valider un appariement.

5. Conclusion et perspectives

Dans cet article, nous avons montré comment réconcilier les données liées à des événements dans le web de données. Nous avons présenté le jeu de données EventMedia composé de descriptions d'événements représentées avec l'ontologie LODE et de descriptions de média représentées avec l'ontologie média du W3C. Nous avons enrichi ce jeu de données à l'aide d'appariements avec d'autres jeux de données et nous avons mis en avant les problèmes rencontrés tels que la nécessité de construire de nouvelles mesures de similarités dédiées ou encore le problème du passage à l'échelle pour apparier de grandes quantités de données. Les appariements entre DBpedia et Freebase étant les mêmes, nous ne les avons pas présentés. L'ensemble des alignements est disponible à `http://www.eurecom.fr/~troncy/ic2011/`.

Notre objectif final est de fournir une interface web permettant à des utilisateurs de découvrir ou de revivre des événements à partir de médias. L'enrichissement sémantique des descriptions a été perçu par les utilisateurs comme un moyen de répondre au problème de la qualité et de la complétude des données que l'on peut trouver dans les annuaires d'événements.

Acknowledgments

Les recherches présentées dans cet article ont été partiellement financé par les projets AAL-2009-2-049 "Adaptable Ambient Living Assistant" (ALIAS) et ANR-2010-CORD-09-02 "Datalift".

Références

CYGANIAK R. & JENTZSCH A. (2010). Linking Open Data cloud diagram. LOD Community. (http://lod-cloud.net/).

EUZENAT J. & SHVAIKO P. (2007). *Ontology Matching*. Database Management & Information Retrieval. Springer-Verlag.

FIALHO A., TRONCY R., HARDMAN L., SAATHOFF C. & SCHERP A. (2010). What's on this evening ? Designing User Support for Event-based Annotation and Exploration of Media. In *1st International Workshop on EVENTS - Recognising and tracking events on the Web and in real life*, p. 40–54, Athens, Greece.

HOBBS J. & PAN F. (2006). Time Ontology in OWL. W3C Working Draft. http://www.w3.org/TR/owl-time.

RAIMOND Y., ABDALLAH S., SANDLER M. & GIASSON F. (2007). The Music Ontology. In *8th International Conference on Music Information Retrieval (ISMIR'07)*, Vienna, Austria.

TRONCY R., FIALHO A., HARDMAN L. & SAATHOFF C. (2010a). Experiencing Events through User-Generated Media. In *1st International Workshop on Consuming Linked Data (COLD'10)*, Shanghai, China.

TRONCY R., MALOCHA B. & FIALHO A. (2010b). Linking Events with Media. In *6th International Conference on Semantic Systems (I-SEMANTICS'10)*, Graz, Austria.

TRONCY R., SHAW R. & HARDMAN L. (2010c). LODE : une ontologie pour représenter des événements dans le web de données. In *21st Journées d'Ingénierie des Connaissances (IC'10)*, p. 69–80, Nîmes, France.

VAN HAGE W., MALAISÉ V., DE VRIES G., SCHREIBER G. & VAN SOMEREN M. (2009). Combining Ship Trajectories and Semantics with the Simple Event Model (SEM). In *1st ACM International Workshop on Events in Multimedia (EiMM'09)*, Beijing, China.

VATANT B. & ROZAT L. (2011). VOAF (Vocabularies of a Friend) vocabulary. Mondeca. (http://www.mondeca.com/foaf/voaf-doc.html).

VOLZ J., BIZER C., GAEDKE M. & KOBILAROV G. (2009a). Discovering and Maintaining Links on the Web of Data. In *2nd Workshop on Linked Data on the Web (LDOW'09)*, Madrid, Spain.

VOLZ J., BIZER C., GAEDKE M. & KOBILAROV G. (2009b). Discovering and Maintaining Links on the Web of Data. In *8th International Semantic Web Conference (ISWC'09)*, p. 650–665, Chantilly, VA, USA.