Terminologie & Ontologie : Théories et applications

Actes de la conférence

TOTh 2010

Annecy – 3 & 4 juin 2010

avec le soutien de :

- Ministère de la Culture et de la Communication, Délégation Générale à la Langue Française et aux Langues de France
- Association Européenne de Terminologie
- Société française de terminologie
- Ecole d'ingénieurs Polytech'Savoie – Université de Savoie
- Université de Sorbonne nouvelle
- Association EGC (Extraction et Gestion des Connaissances)
- ISKO (International Society for Knowledge Organization) France

Institut Porphyre
Savoir et Connaissance

`http://www.porphyre.org`

Comité scientifique

Président du Comité Scientifique : Christophe Roche

Comité de pilotage

Loïc Depecker	Professeur, Université de Sorbonne nouvelle
André Manificat	Directeur, GRETh
Christophe Roche	Professeur, Université de Savoie
Philippe Thoiron	Professeur émérite, Université de Lyon II

Comité de programme

Bruno de Bessé	Professeur, Université de Genève
Franco Bertaccini	Professeur, Université de Bologne
Gerhard Budin	Professeur, Université de Vienne
Marc van Campenhoudt	Professeur, Termisti, ISTI, Bruxelles
Danielle Candel	CNRS, Université Paris Diderot
Stéphane Chaudiron	Professeur, Université de Lille 3
Rute Costa	Professeur, Universidade Nova de Lisboa
Luc Damas	MCF, Université de Savoie
Sylvie Després	Professeur, Université Paris 13
François Gaudin	Professeur, Université de Rouen
Anne-Marie Gendron	Chancellerie fédérale suisse, Section terminologie
Jean-Yves Gresser	Ancien Directeur à la Banque de France
Ollivier Haemmerlé	Professeur, Université de Toulouse
Michèle Hudon	Professeur, Université de Montréal
John Humbley	Professeur, Université Paris 7
Michel Ida	Directeur MINATEC, CEA
Hendrik Kockaert	Professeur, Lessius Hogeschool (Anvers)
Michel Léonard	Professeur, Université de Genève
Pierre Lerat	Professeur honoraire, Equipe Condillac
Widad Mustafa	Professeur, Université de Lille 3
Fidelma Ní Ghallchobhair	Foras na Gaeilge (The Irish-Language Body)
Henrik Nilsson	Terminologicentrum TNC, Suède
Jean Quirion	Professeur, Université d'Ottawa
Renato Reinau	Suva, Lucerne
François Rousselot	MCF, Université de Strasbourg
Gérard Sabah	CNRS, Orsay
Michel Simonet	CNRS, Grenoble
Marcus Spies	Professeur, Université de Munich
Dardo de Vecchi	Professeur associé, Euromed-Management

Comité d'organisation :

Responsable : Luc Damas

Samia Chouder, Joëlle Pellet

Avant propos

Cette année la conférence a été précédée d'une journée de formation consacrée à la terminologie et l'ontologie, à leurs liens et leurs apports mutuels. L'intérêt qu'a suscité cette journée nous amènera certainement à réitérer l'opération les années suivantes.

Le succès de la conférence d'ouverture de notre collègue Frédéric Nef, portant sur l'ontologie prise dans sa dimension philosophique, a montré, s'il en était encore besoin, la richesse d'une approche pluridisciplinaire.

Animées par différents présidents, les sessions ont alterné présentations théoriques et démonstrations de systèmes, offrant ainsi l'opportunité à plusieurs industriels de nous parler de leurs projets. L'éventail des sujets abordés, à travers les quatorze présentations retenues (incluant la conférence d'ouverture) réparties sur deux jours, illustre la richesse mais aussi la vitalité de notre communauté : aide à la traduction, thésaurus multilingue, phraséologie, entité nommé, recherche d'information, etc. L' « actualité » n'a pas été oubliée à travers une ontologie des risques financiers.

Enfin, les Conférences TOTh sont devenues internationales à partir de cette année avec le français et l'anglais comme langues officielles. Le comité de programme s'est ouvert à de nouveaux membres portant à dix le nombre de pays représentés et à plus de 40% le nombre de personnalités étrangères. Gageons que cette ouverture sera prometteuse.

Christophe Roche

Président du Comité Scientifique

Table des matières

CONFERENCE INVITEE

ARTICLES

DEMONSTRATIONS

Modélisation des dénominations ontologiques

Benjamin Diemert, Marie-Hélène Abel, Claude Moulin

Résumé : L'échange de l'information numérique et sa réutilisation dans de nouveaux contextes pose un problème d'écart des représentations entre le producteur et le récepteur. La production télévisuelle se tourne de plus en plus vers des modes de production collaboratifs incluant des amateurs ne maîtrisant pas le jargon audiovisuel. Dans cet article, nous introduisons un modèle conceptuel capable d'articuler différentes dénominations autour d'un même élément ontologique (concept, instance, propriété). Nous définissons notre motivation à partir d'exemples d'échange d'information entre professionnel et amateur puis montrons comment le même problème s'applique également à la recherche multilingue. Nous rappelons l'apport des normes OAIS, FRBRoo et du modèle AXIS à la conceptualisation de l'échange d'information et les manques concernant la modélisation des écarts d'usages de la langue entre différentes communautés. Alors que ces modèles se concentrent sur l'échange de données, nous proposons de décrire et relier l'information, l'utilisateur et son contexte d'action. A partir d'une description d'un utilisateur et de son contexte, nous pouvons alors sélectionner la dénomination qui sera la plus adaptée. La définition de notre modèle permettra de mettre en évidence la généralisation de ce principe à des documents. Nous finissons par discuter nos choix de représentation en OWL.

Mots-clés : Représentation des connaissances, terminologie, archivage numérique, dénomination

1. Introduction

La numérisation massive des contenus et leur mise en réseau ont considérablement augmentées la quantité de documents et de connaissances disponibles dans les organisations. Cependant, si la numérisation doit rendre les contenus plus manipulables et la mise en réseau les rendre plus accessibles, aucun de ces deux procédés n'a pour autant facilité leur exploitation (Abel, 2009).

Le milieu de la production télévisuelle ne fait pas exception. L'exploitation des contenus y est d'autant plus ardue que le processus de création est fortement segmenté. De la définition de l'intention éditoriale à la production en passant par le tournage de scènes et leur montage, chaque étape fait appel à des métiers très spécialisés et donc à des savoir-faire très différents. Remarquons que le matériel et les documents produits à une étape sont réutilisés lors des étapes suivantes.

Par ailleurs, les systèmes informatiques qui ont été développés jusque là répondent à des besoins très divers et se concentrent essentiellement sur le traitement du contenu lui-même. Ils négligent la préservation des informations nécessaires au bon déroulement de la production, ainsi qu'à l'archivage des contenus en vue d'une réutilisation ultérieure. En particulier, manquent les métadonnées décrivant la structure du contenu, le contexte de production, les intentions éditoriales ou en encore les droits concernant les personnes filmées.

Le peu de descriptions présentes est produit de diverses manières. Certaines descriptions sont générées automatiquement par les systèmes de la chaîne de production (coordonnées GPS, caractéristiques de l'objectif d'une caméra). D'autres sont extraites par analyse automatique du contenu (détection de changements de plan, retranscription écrite d'un dialogue par reconnaissance vocale). Certaines enfin sont renseignées manuellement par les différentes communautés d'acteurs de la chaîne de production. Actuellement, ces descriptions sont perdues ou cloisonnées dans des formats ou des schémas de description propriétaires (Gervais et al., 2006, p.7). Le manque d'intégration des systèmes amène à une déperdition de l'information au fur et à mesure de l'avancement dans la chaîne de production (Delvin, 2002).

Il apparaît nécessaire de concevoir un système qui permette à chaque étape d'encapsuler dans une entité commune à la fois le contenu et les descriptions qui seront nécessaires pour une exploitation future, soit durant une étape ultérieure du même processus de production, soit lors d'une recherche dans une archive d'émissions.

Notons que la diversité des communautés présentes amène une diversité de vocabulaires de description même si elles partagent les mêmes concepts. Ces écarts d'usages de la langue posent un problème d'articulation des dénominations (signifiant) au sens des terminologies (Roche, 2005).

Il est donc indispensable de construire un modèle capable d'intégrer ces vocabulaires quel que soit la syntaxe ou le jargon employé. Une conceptualisation avancée, sous forme d'ontologies par exemple, pourra convenir à condition qu'elle prenne en compte les différents contextes d'usages de ces vocabulaires. Le cas du multilingue sera traité comme un cas à part de conceptualisation plus générique.

Dans cet article, nous introduisons les avantages d'une modélisation des dénominations pour l'échange d'information entre communautés ne partageant pas le même jargon ainsi que pour la recherche multilingue (Section 2). Nous présentons ensuite les apports et les manques des modèles dont nous nous inspirons pour notre travail (Section 3). Nous définissons ensuite le modèle et ses principes et reprenons les exemples précédents pour illustrer son utilisation (Section 4). Nous finissons par discuter certains points de notre représentation OWL du modèle (Section 5).

2. Motivations

Ce travail s'effectue dans le cadre du projet MediaMap[1] qui utilise les technologies du Web Sémantique pour favoriser la production collaborative et la réutilisation de contenu audiovisuel. Le consortium comporte notamment les chaînes de télévision publiques belges[2] qui contribuent à définir les besoins du projet. De ce fait, la problématique du multilinguisme est primordiale pour notre approche.

De plus, l'évolution du marché pousse fortement ces organisations à s'orienter vers des modes de production collaborative dans lesquels les sources de contenu se diversifient (tournage, archivage interne, achat). La description, la recherche, l'intégration de contenus provenant de partenaires différents et même d'amateurs deviennent des enjeux cruciaux. L'échange de contenu et de documents ne se fait plus seulement entre différents métiers d'une même organisation, mais entre différents métiers de différentes organisations et

[1] http://www.mediamapproject.org/

[2] La Radio Télévision Flamande (VRT), la Radio Télévision Belge Francophone (RTBF).

même avec des amateurs. Ainsi, l'indexation au cours du processus de production devient nécessaire pour supporter ces échanges.

2.1. Collaboration entre amateur et professionnels

L'appel à la contribution amateur pour la production de contenu professionnel pose en particulier un problème de qualité du contenu. Une solution est de guider l'amateur dans sa tâche à partir d'indications de tournage écrites par un réalisateur dans le jargon de l'audiovisuel. Ces indications sont transmises et traduites en termes compréhensibles par l'amateur. Professionnels et amateurs forment deux communautés distinctes, la première produit l'information, la seconde y accède en vue d'effectuer une tâche. Un exemple d'indication serait un cadrage particulier pour une interview. Voici différentes valeurs de plan utilisées classiquement (voir FIG.1).

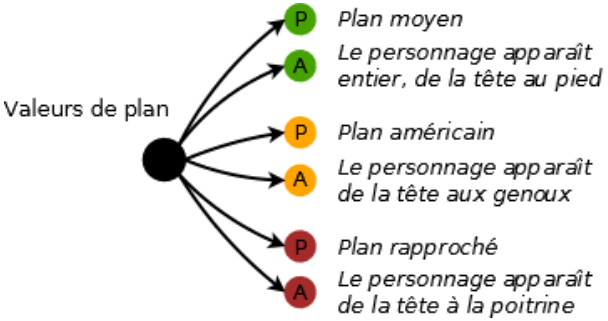

FIG.1 – Valeurs de plan : dénomination dans le jargon Professionnel (P) et pour des Amateurs (A)

Le même concept s'exprime de deux manières, la première (P) est utilisée par les professionnels, la seconde (A) par les amateurs. La distinction entre les deux dénominations à disposition se fait sur le critère de la connaissance du jargon de l'audiovisuel qui définit une communauté par rapport à l'autre.

2.2. Recherche d'information multilingue

De la même manière, on peut considérer le problème du multilinguisme (Ghebghoub, 2009), (Moulin, 2009) en caractérisant les variations orthographiques d'un même mot ou d'une même instance. Par exemple, un nom propre peut avoir différentes orthographes reconnues et chacune sert à identifier la même personne. Le nom du compositeur du Lac des Cygnes possède des dizaines de variantes dont voici trois exemples : Pyotr Ilyich Tchaikovsky (orthographe anglaise); Piotr Ilitch Tchaïkovski (orthographe française); (orthographe russe, voir FIG.1). Imaginons un utilisateur français

lançant une requête sur le compositeur. Du fait de l'orthographe utilisée il risque de ne pas voir les résultats indexés en anglais ou en russe (sans compter les autres variantes) même s'il parle ces langues.

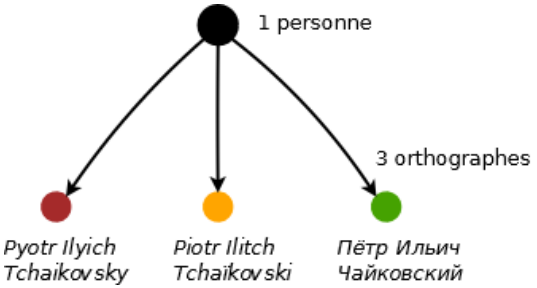

FIG.2 – Différentes orthographes pour le nom du compositeur du Lac des Cygnes

Il apparaît important d'identifier ce qui permet de distinguer ces trois orthographes. Si les deux premières partagent le même alphabet latin, la dernière version est écrite en cyrillique. Si l'on veut différencier les deux premières versions, il faut s'intéresser à la langue. Dans d'autres cas il faut aussi s'intéresser à la localisation géographique (orthographe française différente de l'orthographe québécoise, orthographe anglaise différente de l'américaine etc.).

3. Travaux liés

Parmi les travaux que nous avons examinés, le modèle « Acquisition, eXchange, Indexing and Structuring » (Axis) nous offre la meilleure base conceptuelle pour décrire la production de contenu audiovisuel.

Axis est un modèle et une architecture développés dans le cadre du projet Memories[3] pour traiter le problème de l'échange et l'archivage de ressources média. Axis se veut une implémentation de la norme « Open Archive Information System » (OAIS) qui pose un certain nombre de principes pour l'archivage numérique. Axis se sert de ces principes d'archivage pour résoudre les problèmes d'échanges de ressources entre systèmes différents.

[3] http://www.memories-project.eu/

3.1. L'archivage numérique et l'échange d'information

L'archivage peut être défini dans une perspective documentaire comme une activité intentionnelle de constitution et préservation d'une mémoire. Sa réalisation comprend deux activités indissociables : (1) la conservation physique des documents, (2) la préservation de la tradition de lecture liée aux documents qui assure ainsi leur compréhension (Bachimont, 2000). Ces deux activités sont couvertes à différents degrés par la norme OAIS. Nous présentons d'abord la manière dont OAIS pose le problème de l'archivage puis explicitons l'organisation des échanges d'information.

a) L'archivage comme service d'échange

La norme OAIS définit l'archivage numérique comme un problème d'échange de données entre une communauté productrice et une communauté demandeuse ou réceptrice (CCSDS, 2002). Elle formule des recommandations pour la mise en place d'une organisation rendant un service traitant ce problème (voir FIG.3).

L'enjeu d'un tel service est avant tout d'anticiper les changements de technologies de représentation de l'information dans l'un et l'autre monde. OAIS recommande de conserver certaines informations pour assurer la transformation des données d'une représentation à l'autre, suivant les apports des producteurs et les besoins des consommateurs. Il s'agit donc :

- de préserver les données de toute altération matérielle
- de documenter les données afin de permettre leur utilisation et leur compréhension par la communauté réceptrice.

Notons que cet objectif correspond à une certaine interprétation de la définition de (*Bachimont* 2000). Il faut comprendre ici que la tradition de lecture n'est pas seulement considérée comme la possibilité de voir le document (accès à une forme matérielle de restitution) mais également de le comprendre (accès à une forme sémiotique intelligible) et de l'exploiter pour son usage propre.

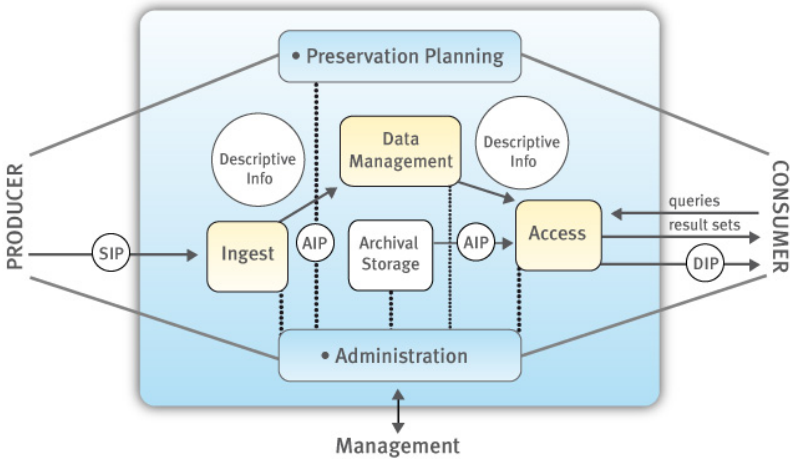

FIG.3 – Un système d'archivage OAIS, organisation fonctionnelle et flux de données

Pour jouer ce rôle d'intermédiaire entre deux mondes et gérer les flux de données la norme pose ainsi une organisation fonctionnelle des échanges et de l'archivage ainsi qu'un format d'encapsulation des données, les Information Package (IP).

b) Gérer l'échange de données

Nous détaillons l'organisation fonctionnelle proposée par OAIS (voir FIG. 3). D'abord l'*Ingest* réceptionne les informations fournit par le *Producer*. Il les transforme en « Archival Information Package » (AIP) répondant aux standards définit par l'*Administration* est constitué à partir des « Submission IP » (SIP).

Le contenu de l'AIP est pris en charge par l'*Archival Storage*, sa description par le *Data Management*. Le contenu et sa description sont donc stockés séparément du fait de besoins opérationnels différents. En effet, la recherche d'information se fait à partir des descriptions (elles sont donc fréquemment consultées) alors que l'accès au contenu se fait uniquement au moment de la livraison ou lors de contrôle d'intégrité.

L'*Access* utilise ces descriptions pour répondre aux recherches de la communauté de consommateurs. Au moment de la livraison, on prépare un « Dissemination IP » (DIP) à partir de l'AIP suivant les exigences des consommateurs. Ainsi, chacun des IP répond à des exigences différentes (la livraison, l'archivage, la diffusion) ce qui nécessite parfois des transformations

(formats) ou des réarrangements (composition). Ces modifications des IP ne sauraient se faire sans un modèle d'information qui structure et documente le contenu.

c) Documenter les données

Le modèle d'information préconisé dans OAIS distingue quatre types d'information :

- *Description du contenu* : Le contenu à archiver doit être accompagné d'une description qui permette de traduire les séquences de bits en information intelligible pour la communauté réceptrice. La description porte sur les structures de données, sur ce qu'elles représentent et les connaissances nécessaires pour interpréter correctement le contenu. Pour un fichier PDF contenant une partition musicale, il faut d'une part définir le format PDF pour accéder à la partition (accès à une forme matérielle) et d'autre part définir la notation musicale (accès à une forme intelligible).
- *Information pour la préservation* : La préservation du contenu passe avant tout par une identification robuste de tout ou partie du contenu ainsi qu'une description de son cycle de vie. En partant du contexte de production qui permet d'éclairer les intentions du producteur, un historique des transformations doit être tenu. Des indicateurs de qualité doivent être définis pour tester l'intégrité des données au bit près si nécessaire (identification, gestion du cycle de vie).
- *Description du contenu d'un IP* : L'information décrivant le tout et les parties du contenu à archiver en vue d'une recherche d'information.
- *Description de l'organisation d'un IP* : L'information indiquant le lien entre chaque partie de l'IP et un emplacement sur un support physique.

d) Manques

OAIS permet d'identifier des exigences d'organisation fonctionnelle et de documentation des données. Cependant, aucun modèle de données n'est proposé. De plus, la norme n'aborde pas la question des différences de pratiques des communautés que sous l'angle de la représentation de l'information. La question des usages de la langue et de la représentation des connaissances des communautés est simplement mentionnée comme étant à traiter.

3.2. FRBRoo

« Functional Requirements for Bibliographic Records » (FRBRoo) est un modèle conceptuel développé pour faciliter l'échange d'information entre les librairies numériques et les musées (CIDOC, 2009). Il permet de représenter les personnes participant aux différentes étapes de construction d'un objet culturel, depuis l'idée jusqu'à la réalisation matérielle. Chaque objet possède trois types de caractéristiques :

- Les idées abstraites (Work) n'ayant pas pris corps dans une matérialité externe à un sujet (une mélodie ou une histoire).
- Les différentes expressions (Expression) possibles pour une idée (une nouvelle écrite, ses traductions, une adaptation de la nouvelle en scénario etc.).
- Les porteurs physique d'information (Information Carrier) qui portent les expressions (livre, partition, cd-rom etc.). Dans ce cas, le modèle distingue entre l'original (Manifestation Singleton), des copies manufacturées (Item) issues d'un modèle de publication (Manifestation Product Type).

3.3. Axis

Axis reprend les exigences d'OAIS tout en les adaptant pour l'échange d'information et de ressources médias entre organisations. Les conséquences sont multiples. D'abord l'architecture du système se doit d'être flexible et distribuée. Chaque organisation peut remplir une ou plusieurs fonctions, parfois de manière redondante (*Ingest*, *Access*).

Ensuite, Axis propose de n'utiliser qu'un seul type de paquet d'information, l'« Autonomous eXchange Entity » (AXE), pour permettre l'échange. La fusion des paquets doit répondre aux exigences de l'archivage et du même coup à celui de l'interopérabilité entre systèmes.

La construction du modèle Axis s'inspire également du modèle conceptuel FRBRoo pour décrire les ressources médias comme des objets culturels. Le modèle Axis se réapproprie ces concepts en les considérant comme trois niveaux distincts ; le niveau conceptuel, le niveau des signes, le niveau des supports physique. Expression et Manifestation deviennent ainsi des relations entre objets distincts. Ainsi pour une idée, on peut associer différentes expressions sémiotiques, chacune pouvant être inscrites sur un ou plusieurs supports.

Au niveau des signes, Axis utilise le principe des thésaurus de manière similaire au standard du W3C « Simple Knowledge Organization System » (SKOS). Une approche de représentation de type thésaurus permet en effet de relier chacune des dénominations possibles (alternatives) à une dénomination de référence. SKOS par exemple permet de les grouper en schémas et de représenter les relations entre dénominations (spécificité, généralité, utilisé pour). La modélisation en OWL Full de SKOS permet de représenter les thésaurus existants et de les intégrer au Web Sémantique avec toutes les possibilités d'alignement qui s'en suivent (Summers et al., 2008).

Cependant cette approche ne permet pas de caractériser les contextes d'usages des dénominations. Le seul attribut disponible est « xml:lang » qui permet d'ajouter le code d'une langue ce qui permet de couvrir le cas du multilingue, mais pas du tout le cas des jargons métiers (voir Section 2). Une solution consiste à avoir un schéma par contexte d'usage et de les relier manuellement.

4. Modélisation

Nous présentons d'abord les principes généraux qui fondent notre modélisation (Section 4.1), puis nous montrerons la modélisation des exemples précédemment décrits (Section 4.2). Nous finirons par une définition détaillée des concepts et relations de notre modèle (Section 4.3).

4.1. Principes

Nous avons trouvé dans Axis un modèle générique qui partage une bonne partie des problématiques du projet MediaMap. Les concepts et les principes d'Axis servent de fondation à notre travail qui s'étend à la modélisation des contextes d'usages des dénominations (signifiant) attachées à un même concept, une même propriété ou une même instance. Notre approche tire partie de façon intéressante des modèles existants.

Notre modélisation permet de faire un choix informé parmi différentes dénominations à disposition. Elle intègre la description de la situation de production et d'accès à l'information en prenant en compte l'utilisateur et son contexte. Ces éléments contextuels servent d'une part à identifier des usages de la langue, et d'autre part à décrire un utilisateur et le relier à des usages identifiés. De cette manière, on peut proposer automatiquement une dénomination adaptée à un utilisateur connu.

Le modèle se divise en trois parties (voir FIG. 4). La première définit les capacités de l'utilisateur (Modèle Utilisateur). La deuxième regroupe les dénominations selon leurs caractéristiques (Modèle d'Information). La dernière est composée d'éléments généraux qui permettent de spécifier et de relier le contexte d'usage des dénominations au contexte d'action d'un utilisateur (Modèle du Contexte).

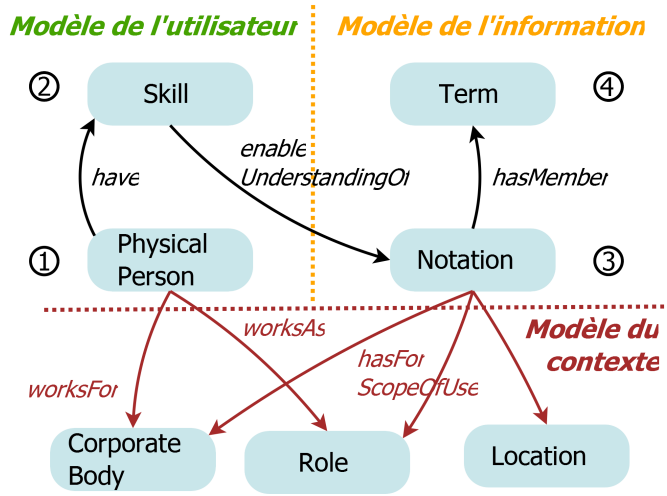

FIG. 4 – Concepts principaux et Relations entre Modèle Utilisateur ; Modèle d'Information ; Modèle du Contexte

Ainsi, la sélection d'une dénomination (instance de Term) se fait d'abord en fonction des capacités (instance de Skill) d'un utilisateur (instance de PhysicalPerson). Ces capacités permettent de comprendre un jargon métier, une langue ou un codage (instance d'une spécialisation de Notation). D'autres éléments contextuels peuvent permettre de sélectionner plus finement une dénomination et d'identifier une communauté d'usage de la langue. Par exemple l'appartenance à une organisation (instance de CorporateBody) qui a défini un thésaurus métier. Le métier de l'utilisateur (instance de Role) peut également apporter des indications précieuses sur le vocabulaire et les codes utilisées. Le lieu de vie ou le lieu de travail (Location) permet également d'identifier l'appartenance à une communauté linguistique (alphabet, langue).

Finalement, il faut préciser que le modèle peut s'étendre à la caractérisation de tous types de support d'information (instance de Document).

4.2. Modélisation des exemples

Nous présentons une vue simplifiée de la modélisation des exemples présentés en Section 2. Leurs représentations en OWL sont détaillées en Section 5. D'abord, les dénominations de Valeur de Plan pour amateurs et professionnels. La FIG. 5 présente une vue simplifiée du réseau des relations entre deux utilisateurs et deux dénominations différentes pour désigner la même valeur de plan.

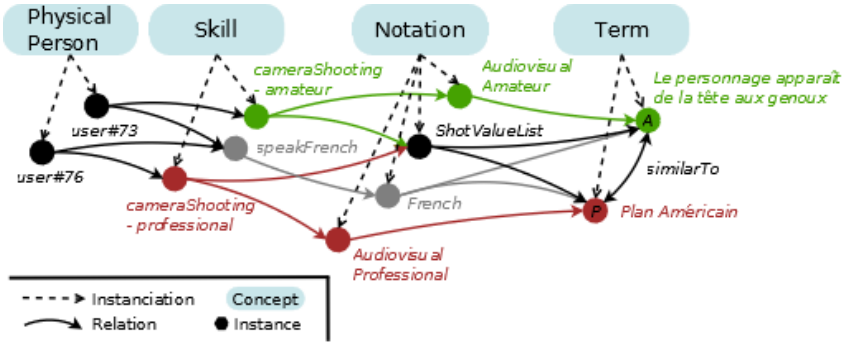

FIG. 5 – Vue simplifiée du réseau de relations entre deux utilisateurs et deux dénominations d'une même valeur de plan

L'exemple considère deux caméramans francophones, l'un amateur, l'autre dont c'est le métier. Tous deux doivent donc comprendre les indications de tournage qu'on leur donne, dont la liste des valeurs de plan. On les distingue par leur niveau de compétence (amateur, professionnel) qui se traduit par l'accès à deux vocabulaires différents (instance de Notation) pour désigner les valeurs de plan. Chaque valeur de plan de la liste est exprimée par deux dénominations distinctes (instance de Term) appartenant à l'un ou l'autre vocabulaire. Comme il n'y a pas un concept par valeur de plan, mais juste des dénominations, une relation de similarité est nécessaire pour regrouper celles qui correspondent à la même valeur de plan.

L'exemple de Tchaïkovski (voir FIG. 6) permet de montrer un exemple de recherche d'information multilingue. Supposons que la requête d'un utilisateur soit exprimée en français et aboutissent à retrouver le compositeur (instance de Personne). Dans ce cas, tous les documents (A, B, C) relatifs au compositeur seront retrouvés. Grâce à la description des capacités linguistiques de l'utilisateur, le système pourra mettre en avant ceux qui lui sont compréhensibles.

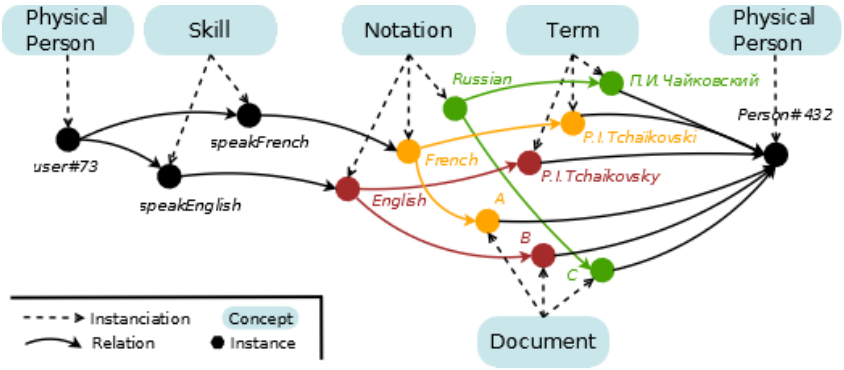

FIG. 6 – Recherche multilingue et sélection de document

On peut complexifier l'exemple en rajoutant à la description des documents l'encodage des caractères (ISO 8859-1 pour l'alphabet latin, ISO 8859-5 pour le cyrillique). Cet ajout peut servir lorsqu'on décrit également le terminal d'accès utilisé par un utilisateur et que l'on renseigne ses capacités à lire tel ou tel encodage.

4.3. Description de la conceptualisation

Nous définissons les concepts et les relations des différentes parties de notre modèle. Nous finissons par présenter les relations qui permettent de relier un utilisateur à des dénominations ou d'autres types d'information.

a) Concepts et Relations du Modèle d'information

Un terme (instance de la classe **Term**) encapsule la valeur d'une propriété qui dans une modélisation commune serait généralement de type chaîne de caractères. Nous lui adjoignons un ensemble de caractéristiques qui précisent son contexte d'utilisation. Le véritable type de la valeur peut en fait être quelconque : une chaîne, mais aussi un nombre, une date, etc.

Un document (instance de la classe **Document**) désigne un support d'information en relation avec un concept. Les concepts sont des entités abstraites qui ne sont accessibles qu'à travers des manifestations physiques que sont les documents. Ainsi contrairement à FRBRoo, une manifestation est une relation (**isAManifestationOf**) entre un document et une entité.

Une notation (instance de la classe **Notation**) dénote un ensemble de termes partageant des caractéristiques communes. La classe Notation possède un ensemble de spécialisations que nous détaillons ci-dessous.

- Un langage naturel (instance de la classe **NaturalLanguage**) désigne l'appartenance d'un terme à un langage naturel, (français, anglais, flamand etc.). Le code d'un pays modélise les variations géographiques d'une langue.
  Par exemple l'anglais est parlé dans différentes régions du monde et des distinctions existent entre l'anglais du Royaume-Uni et celui des Etats-Unis d'Amérique.

- Un schéma d'encodage syntaxique (instance de la classe **SyntaxEncodingScheme**) désigne un codage particulier d'octets permettant de représenter une valeur.
  Ce peut être un codage de caractères (UTF-8), un encodage de type de données (xsd:string), un codage avec une syntaxe complexe tel un format de fichier (jpg, pdf) ou des formats de dates (ISO 8601 : 1988 (E)[4]).

- Un schéma d'encodage de termes d'un vocabulaire (instance de la classe **VocabularyEncodingScheme**) désigne un ensemble d'appartenance de termes tel qu'un thésaurus métier. Cela permet plus généralement de modéliser une convention de nommage dans un contexte métier, organisationnel ou géographique. Un schéma de description de métadonnées tel que Dublin Core[5] est un exemple.

- Une liste d'autorité (instance de la classe **AuthorityList**) désigne un ensemble fini de termes dont les valeurs sont arbitrairement choisies.
  Ainsi, la norme « ISO 639-2[6] » est une liste de codes à trois lettres dénotant le nom des langues. Par exemple, « eng » pour l'anglais.

Un **terme** est membre (**isMemberOf**) d'une notation lorsqu'il remplit les caractéristiques qui la définissent. L'appartenance à une notation n'est pas exclusive ce qui permet de raffiner les descriptions jusqu'au niveau de détails souhaité. Cette relation se spécialise par la relation d'appartenance à une langue (**inLanguage**).

[4] Les formats de date conseillés par le W3C, http://www.w3.org/TR/NOTE-datetime

[5] http://dublincore.org/documents/dcmi-terms/

[6] http://www.loc.gov/standards/iso639-2/

Une **notation** se conforme à (**conformsTo**) à un schéma d'encodage syntaxique (SES) lorsqu'elle utilise ses règles de codage. Une notation peut se conformer à plusieurs schémas ce qui permet là encore de détailler la description à souhait.

b) Concepts et Relations du Modèle utilisateur

Un agent (instance de la classe **Agent**) désigne une entité agissante. Cette classe possède trois spécialisations :

- Une personne (instance de la classe **PhysicalPerson**) désigne un individu, « Tchaïkovski » par exemple.
- Une personne morale (instance de la classe **CorporateBody**) désigne une organisation composée d'autres agents, le « Minneapolis Symphony Orchestra ».
- Un système (instance de la classe **Proxy**) désigne un système agissant pour le compte d'un autre agent, un serveur web.

Une compétence (instance de la classe **Skill**) désigne la capacité d'un agent à effectuer une tâche ou à comprendre une information en rapport avec cette tâche (indications de réglage, mode d'emploi pour un nouvel instrument etc.). Cette classe possède deux spécialisations :

- Une compétence linguistique (instance de la classe **LinguisticSkill**) désigne la capacité d'un agent à comprendre une langue.
- Une compétence professionnelle (instance de la classe **BusinessSkill**) désigne la capacité d'un agent à effectuer des tâches professionnelles et à comprendre des notations métiers.
  « SavoirCadrer » désigne ainsi la compétence d'une personne à régler une caméra en fonction d'une indication de valeur de plan telles que plan américain, gros plan, plan d'ensemble. Ces différentes valeurs étant des termes regroupés dans une liste d'autorité.

Une personne est définit comme compétente (**isProficient**) lorsqu'on lui reconnaît cette capacité.

c) Concepts et Relations du Modèle du Contexte

Les compétences permettent de caractériser des métiers (instance de la classe **Role**). Un rôle désigne alors les droits et les responsabilités que l'on attribue à une personne ou une organisation dans un projet ou un processus.

Une personne travaille (**workFor**) pour une organisation lorsque cette dernière l'emploie contractuellement. De même, un emploi implique qu'on attribue (**workAs**) une position ou un métier à cette personne.

Un lieu (instance de **Location**) désigne une entité géographique ou politique. Location est un concept général qui peut être spécialisé suivant les besoins.

d) Relations entre parties du modèle

Il y a trois relations très importantes dans le modèle. Celle qui permet de relier l'utilisateur à des notations, celle qui relie les notations à des éléments contextuels et enfin celle qui permet de relier un terme à un élément ontologique :

- Une compétence désigne la capacité d'un agent à comprendre une notation (**enablesUnderstandingOf**)
- Une notation a pour contexte d'usage (**hasForScopeOfUse**) une organisation (instance de CorporateBody), un lieu (instance de Location) ou un type de métier (instance de Role).
- La relation de nommage (**isNamedBy**) permet d'attacher un terme à n'importe quel autre élément du modèle. Ainsi, on généralise les mécanismes de caractérisations des dénominations aux concepts, instances, propriétés.

5. Représentation OWL

La représentation de notre modèle sous forme d'ontologie nous semble profitable sur plusieurs points. Tout d'abord le fait d'expliciter la sémantique des concepts utilisés permet de faciliter la compréhension du modèle, donc son appropriation par les utilisateurs et sa maintenance pour les concepteurs. De plus, le fait de modéliser des objets d'un domaine sans concentrer son attention sur une tâche particulière permet d'obtenir des conceptualisations plus flexibles et extensibles (Spyns et al., 2002). Lorsqu'il s'agit ensuite d'effectuer des recherches sur les données produites à partir d'une ontologie, on mesure les bénéfices d'un tel détachement grâce aux extensions de requêtes (Guarino, 1998). Enfin, les capacités de raisonnement et d'intégration de données entre modèles différents sont des avantages particulièrement précieux lorsqu'on traite une problématique d'échange d'information (Garcia et al., 2005).

Les langages de représentation de connaissances telle que RDF et OWL nous offrent une manière d'exprimer notre ontologie sous une forme répandue et exploitable. Nous avons choisi d'utiliser la variante OWL-DL pour s'assurer une expressivité forte sans pour autant sacrifier à la calculabilité. Nous discutons de certains détails à partir des exemples présentées en Section 2 et modélisé en Section 4.2.

5.1. Dénominations pour amateur et professionnel

Voici la modélisation de l'exemple de la liste de Valeur de Plan en notation N3[7]. Une liste d'autorité a pour membre des instances de termes dont certains renvoient à la même valeur (un plan américain). Cependant cette valeur n'est pas représentée en tant que telle, elle n'existe que par les termes qui appartiennent à l'un ou l'autre vocabulaire (amateur, professionnel).

Il faut donc pouvoir d'une part caractériser chaque terme de la liste individuellement (par les relations isMemberOf, inLanguage) et d'autre part identifier la similarité entre deux termes (par la relation similarTo).

```
// ==== NOTATIONS ===== //
mm:shotValueList
    a      mm:AuthorityList ;
    mm:hasMember mm:planAmericain-a , mm:planRapproche , mm:planRapproche-a ,
mm:planAmericain .
mm:Audiovisual-Professional
    a      mm:VocabularyEncodingScheme ;
    mm:hasMember mm:planAmericain , mm:planRapproche .
mm:Audiovisual-Amateur
    a      mm:VocabularyEncodingScheme ;
    mm:hasMember mm:planAmericain-a , mm:planRapproche-a .
mm:French-NL
    a      mm:NaturalLanguage .

// ==== TERM ===== //
mm:planAmericain
    a      mm:Term ;
    mm:inLanguage mm:French-NL ;
    mm:isMemberOf mm:shotValueList , mm:Audiovisual-Professional ;
    mm:similarTo mm:planAmericain-a ;
    mm:value "Plan Américain"^^xsd:string .
```

[7] http://www.w3.org/DesignIssues/Notation3.html

```
mm:planAmericain-a
    a       mm:Term ;
    mm:inLanguage mm:French-NL ;
    mm:isMemberOf mm:shotValueList  , mm:Audiovisual-Amateur ;
    mm:similarTo mm:planAmericain ;
    mm:value "Plan qui fait apparaître le personnage de la tête jusqu'aux genoux."^^xsd:string
.
```

La description des utilisateurs se fait par les compétences et leur niveau. Le niveau de compétence est un attribut qui peut prendre deux valeurs : amateur ou professionnel. Ainsi, chaque compétence peut avoir deux variantes qui renvoient à deux vocabulaires distincts (Audiovisual-Amateur, Audiovisual-Professional). Ces variantes permettent également d'accéder à des notations communes, ici une liste d'autorité de valeur de plan.

```
// ==== PHYSICAL PERSON ===== //
mm:user_73
    a       mm:PhysicalPerson ;
    mm:isProficient mm:cameraShooting-a , mm:speakFrench .
mm:user_76
    a       mm:PhysicalPerson ;
    mm:isProficient mm:cameraShooting-p , mm:speakFrench .

// ==== SKILLS ===== //
mm:speakFrench
    a       mm:LinguisticSkill ;
    mm:enablesUnderstandingOf
            mm:French-NL .

mm:cameraShooting-p
    a       mm:BusinessSkill ;
     mm:enablesUnderstandingOf
            mm:Audiovisual-Professional , mm:shotValueList ;
    mm:SkillLevel "professional"^^xsd:string .
mm:cameraShooting-a
    a       mm:BusinessSkill ;
    mm:enablesUnderstandingOf
            mm:Audiovisual-Amateur , mm:shotValueList;
    mm:SkillLevel "amateur"^^xsd:string .
```

5.2. Dénominations et recherche d'information multilingue

L'exemple avec Tchaïkovski permet de discuter de la représentation en OWL de la relation de nommage (isNamedBy). Il s'agit nécessairement d'une propriété d'annotation (AnnotationProperty) car elle doit pouvoir attacher un terme à n'importe quel élément ontologique (concept, propriété, instance). De ce fait, elle se substitue à l'usage de rdfs:label tout en le généralisant. Notre choix de représentation en OWL-DL implique certaines contraintes[8] qui n'ont pas de conséquences pour l'usage que nous avons choisi. Les triplets RDF que nous construisons sont systématiquement de la forme suivante : « sujet mm:isNamedBy terme » avec comme sujet une classe, une propriété ou un individu et comme objet une instance de Term.

Voici la représentation des trois variantes de termes nommant Tchaïkovski, des documents se rapportant au compositeur (par la relation isAManifestationOf) et des langues dans lesquels le contenu des documents ou les dénominations sont exprimées (par la relation inLanguage).

```
// ==== NOTATIONS ===== //
mm:French-NL
    a      mm:NaturalLanguage .
mm:Russian-NL
    a      mm:NaturalLanguage .
mm:English-NL
    a      mm:NaturalLanguage .

// ==== TERM ===== //
mm:Tchaikovsky-EN-Term
    a      mm:Term ;
    mm:inLanguage mm:English-NL ;
    mm:value "Pyotr Ilyich Tchaikovsky"^^xsd:string .
mm:Tchaikovsky-FR-Term
    a      mm:Term ;
    mm:inLanguage mm:French-NL ;
    mm:value "Piotr Ilitch Tchaïkovski"^^xsd:string .
mm:Tchaikovsky-RU-Term
    a      mm:Term ;
    mm:inLanguage mm:Russian-NL ;
    mm:value "Пётр Ильич Чайковский"^^xsd:string .
```

[8] Dans la variante OWL-DL, il n'est pas possible de définir de domaine, portée et de sous-propriété à une propriété d'annotation.

```
// ==== DOCUMENT ===== //
mm:Document_A
    a      mm:Document ;
    mm:inLanguage mm:French-NL ;
    mm:isAManifestationOf
          mm:Tchaikovsky .
mm:Document_B
    a      mm:Document ;
    mm:inLanguage mm:English-NL ;
    mm:isAManifestationOf
          mm:Tchaikovsky .
mm:Document_C
    a      mm:Document ;
    mm:inLanguage mm:Russian-NL ;
    mm:isAManifestationOf
          mm:Tchaikovsky .
```

Une instance de personne peut se nommer d'autant de manières souhaitées (par la relation isNamedBy). Un utilisateur lançant une requête à partir de l'orthographe française du compositeur aboutit à retrouver la personne et tous les documents associés. Suivant la description des documents et les compétences de l'utilisateur, un système peut mettre en avant les contenus qu'il est le plus susceptible de comprendre (ici les documents en français et en anglais).

```
// ==== PHYSICAL PERSON ===== //
mm:Tchaikovsky
    a      mm:PhysicalPerson ;
    mm:  isNamedBy mm:Tchaikovsky-FR-Term , mm:Tchaikovsky-RU-Term , mm:Tchaikovsky-
EN-Term .
mm:user_73
    a      mm:PhysicalPerson ;
    mm:isProficient mm:speakEnglish , mm:speakFrench .

// ==== SKILLS ===== //
mm:speakFrench
    a      mm:LinguisticSkill ;
    mm:enablesUnderstandingOf
          mm:French-NL .
```

```
mm:speakEnglish
    a      mm:LinguisticSkill ;
    mm:enablesUnderstandingOf
          mm:English-NL .
```

6. Conclusion

Dans cet article nous avons décrit un modèle conceptuel permettant de caractériser la portée d'usage des dénominations se rapportant à tous types d'éléments ontologiques (concept, instance, propriété). Cette caractérisation se fait sur la base d'éléments contextuels (métier, organisation, lieu) qui servent de critères communs à une description de l'information et de l'utilisateur voulant y accéder. Ce lien permet de sélectionner la dénomination ou la forme d'information la plus pertinente pour une communauté d'utilisateur donné.

Nous avons d'abord montré la pertinence de cette approche dans le contexte d'une production télévisuelle collaborative où les contributions d'amateurs sont amenées à prendre de l'importance. La capacité à échanger de l'information et à la réutiliser dans de nouveaux contextes devient un enjeu important pour les organisations. Notre étude d'un modèle conceptuel d'archivage numérique permet de poser le problème comme celui d'un échange d'information entre communautés dont les représentations varient. Cependant, les efforts se concentrent sur l'échange matériel des données plutôt que sur la représentation des connaissances et des usages de la langue nécessaires à leur interprétation. L'étude de modèles de représentation des objets culturels et des ressources médias nous a permis de poser une distinction entre concept, signe et porteur physique d'information. Cette distinction nous a servi de base pour le développement de notre modèle qui aboutit à une ontologie en OWL-DL.

Notre modèle permet de sélectionner la forme d'information la plus pertinente pour un utilisateur souhaitant y accéder. La recherche d'information ainsi que la compréhension du contenu s'en trouve facilitée. Nous travaillons actuellement à la modélisation du processus de production télévisuelle pour affiner notre modèle du contexte. L'objectif est de faciliter la communication entre les différents acteurs participant au processus de production en précisant les informations qu'ils produisent et qu'ils échangent.

Bibliographie

Abel M., Leblanc A. (2009) : *Knowledge Sharing via the E-MEMORAe2.0 Platform,* In 6th International Conference on Intellectual Capital, Knowledge Management & Organisational Learning (pp. pp. 10-19). Montreal, Canada.

Bachimont Bruno (2000) : *L'archive numérique: entre authenticité et interpretabilité.* archivistes.qc.ca

Centre Internation de la DOCumenation (CIDOC) (2009) : *FRBR object-oriented definition and mapping to FRBRer*, http://cidoc.ics.forth.gr/frbr_inro.html.

Roche C. (2005) : *Terminologie & Ontologie,* Langages, 157, 1-11.

Consultative Committee for Space Data Systems (CCSDS) (2002) : *Reference Model for an Open Archival Information System*, http://public.ccsds.org/publications/archive/650x0b1.pdf

Devlin B. (2002) : *What is MXF?,* Ebu Technical Review, (July), 1-7.

Garcia R., Celma O. (2005) : *Semantic Integration and Retrieval of Multimedia Metadata,* In 5th International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot'05). Ireland.

Ghebghoub O., Moulin C., Abel M. (2009) *: Création d'une ontologie des ressources numériques, In* 3èmes Journées Francophones sur les Ontologies (JFO) (pp. 65-71). Poitiers, France.

Gouyet J., Gervais J. (2006) : *Gestion des médias numériques*, (INA) Audio-Photo-Video (p. 328). Paris, France: Dunod.

Moulin C., Sugawara K., Fujita S., Wouters L., Manabe Y. (2009) : *Multilingual Collaborative Design Support System,* In 13th International Conference on Computer Supported Cooperative Work in Design (CSCWD 2009) (pp. 312-318). Santiago, Chile.

Summers E., Isaac A., Redding C., Krech D. (2008) : *LCSH, SKOS and Linked Data,* CoRR.

A propos des auteurs

Diemert Benjamin (A)
benjamin.diemert@hds.utc.fr
http://www.hds.utc.fr/~bdiemert/perso/

Abel Marie-Hélène (A)
marie-helene.abel@hds.utc.fr
http://www.hds.utc.fr/~mhabel/perso/

Moulin Claude (A)
claude.moulin@hds.utc.fr
http://www.hds.utc.fr/~cmoulin/perso/

(A) Heudiasyc, UMR 6599
Université de Technologie de Compiègne
Centre de Recherches de Royallieu
BP 20529
60205 COMPIEGNE cedex FRANCE

TOTh 2010. *Actes de la quatrième conférence TOTh - Annecy – 3 & 4 juin 2010*
Editeur : Institut Porphyre, *Savoir et Connaissance*
http://www.porphyre.org
Annecy, 2010
ISBN 978-2-9536168-1-1
EAN 9782953616811