PICS Distributed Integration, Gradening and Investigation for Cultural Knowledge with Streams (#6945)

Janvier 2015 à Décembre 2017

Responsable scientifique : *Duchateau Fabien, Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS), UMR5205, Université Claude Bernard Lyon 1*

RAPPORT D'ACTIVITÉ 2017

A. MOBILITE TRANSNATIONALE

A.1- Organisation de réunions de travail sur la thématique du PICS

Indiquer l'objet de la réunion, date, lieu (laboratoire partenaire ou ville), nombre total de participants, identifier les participants français (nom, qualité, laboratoire de rattachement, durée de la mission).
Vous pouvez également donner ces renseignements sous forme de tableau Excel.

Cette année, un seul séjour dans le laboratoire partenaire, et donc les échanges (visioconférences, emails et espaces collaboratifs) ont été plus fréquents pour vérifier l'état d'avancement du projet et de coordonner les tâches.
Par exemple, la rédaction d'un article journal a nécessité différentes périodes d'intenses échanges entre Joffrey Decourselle (LIRIS, doctorant), Trond Aalberg (NTNU, professeur), Fabien Duchateau (LIRIS, enseignant-chercheur) et Naimdjon Takhirov (Westerdals university of Oslo, chercheur associé) : rédaction de la première version pour fin janvier 2017, seconde version pour fin avril 2017, troisième version pour fin août 2017.
Autre exemple, des réunions mensuelles (visioconférence) ont lieu entre Audun Vennesland (NTNU, doctorant) et Fabien Duchateau (LIRIS, enseignant-chercheur), rejoints par Trond Aalberg (NTNU, professeur) de manière plus ponctuelle. Ces réunions ont pour objectif d'avancer les travaux sur la combinaison d'alignements pour améliorer la tâche d'alignement.
Les deux doctorants (Joffrey Decourselle et Audun Vennesland) échangent également régulièrement, notamment sur la construction de la plateforme.

A.2 - Accueil, dans le laboratoire français, de chercheurs des laboratoires partenaires étrangers

Objet de l'accueil, date, nom du chercheur, qualité, laboratoires d'origine et d'accueil, durée du séjour, si le chercheur a donné un séminaire indiquer le titre
Vous pouvez également donner ces renseignements sous forme de tableau Excel.

Nos partenaires norvégiens ne bénéficiaient pas en 2017 d'un financement spécifique pour des missions au LIRIS. Ils ont à disposition des sources de financement plus générales, par exemple pour favoriser le montage de projets européens ou des fonds récurrents (annuels).
Trond Aalberg avait prévu un séjour au LIRIS fin novembre (réservations effectuées pendant l'été). Ce séjour a dû être annulé car il a été invité comme « keynote speaker » à la conférence Metadata and Semantics Research Conference sur cette même période (http://www.mtsr-conf.org/index.php/keynotes).
Ce séjour est reporté au printemps 2018 en vue de préparer un montage de projet européen.

A.3 - Séjours, dans le laboratoire partenaire étranger, de chercheurs du laboratoire français

Objet du séjour, date, nom du chercheur, qualité, laboratoires d'origine et d'accueil, durée du séjour, si le chercheur a donné un séminaire indiquer le titre

Objet	Date	Chercheur français	Labo accueil	Durée
Rédaction des articles pour JCDL 2017, avancements sur l'enrichissement	14/12/16	Joffrey Decourselle (doctorant, LIRIS)	NTNU	6 jours

Co-rédaction de la seconde version de l'article journal IJDL. Discussions et programmation pour l'utilisation de l'outil FEBRL dans les expérimentations de combinaison d'alignement. Identification d'un sujet et de partenaires pour montage de projet européen	27/06/17	Nicolas Lumineau et Fabien Duchateau (LIRIS, enseignant-chercheurs)	NTNU	8 jours

Pour rappel, le doctorant Joffrey Decourselle était parti en mission à NTNU en décembre 2016, mais une partie de sa mission (frais de séjour) ne pouvait pas être remboursée sur le budget 2016 (et donc remboursée sur le budget 2017, d'où la présence de cette mission dans le tableau).

A.4 – Organisation de conférences, écoles d'été, ateliers etc. par les partenaires du PICS
Objet, date, lieu, organisateur, nombre total de participants, identifier les participants du laboratoire français (nom, qualité, laboratoire de rattachement, durée de la mission)
Vous pouvez également donner ces renseignements sous forme de tableau Excel.

B. TRAVAUX EN COLLABORATION

B.1 – Etat d'avancement du projet scientifique du PICS
5 pages maximum pour l'année en cours ou 15 pages maximum pour les projets arrivant au terme des 3 ans. Le nom des chercheurs impliqués sera précisé

Résumé du projet

Dans le domaine de l'héritage culturel, la diversité des modèles et la dispersion des données conduit à l'obtention de sources de données distribuées, redondantes et incohérentes, ce qui complique grandement les recherches des spécialistes comme du grand public. Le projet DIRICKS s'intéresse à la construction, la maintenance et l'interrogation de bases de connaissances dans un contexte où les données traitées sont hétérogènes, distribuées et peuvent se présenter sous la forme de corpus ou de flux continu d'informations. Les défis scientifiques abordés dans ce projet concernent plusieurs domaines, notamment l'intégration et l'enrichissement de données multi-sources, l'organisation et l'indexation de données sur des réseaux large échelle, le maintien de la cohérence des données, le requêtage distribué couplé à de nouvelles méthodes de recherche exploratoire et, à la distribution des traitements, clé de voûte du passage à l'échelle.

Dans le projet DIRICKS, les tâches étaient réparties en cinq work packages (WP) :
- WP1 (dataset and model)
- WP2 (matching)
- WP3 (gardening)
- WP4 (query processing)
- WP5 (dissemination)

Dans le reste de cette partie nous détaillons les avancées effectuées pour chacun de ces work packages, et nous terminons par une conclusion sur le projet ainsi que les perspectives, tant sur le plan scientifique que sur le plan collaboration.

Pour le work package WP1 (dataset and model) :

Les trois tâches associées à ce work package concernent la modélisation de la base de connaissances thématique (knowledge base meta-modelling), la construction d'un benchmark et l'extraction des entités (entity extraction).

- La première tâche à laquelle nous nous sommes confrontés dans ce work package concerne le méta-modèle de la base de connaissances. C'est une étape cruciale puisque le modèle de la base de connaissances doit être à la fois suffisamment riche pour inclure de nombreux concepts présents dans d'autres sources (enrichissement) et compréhensible par les acteur/rice/s du domaine culturel. Ce domaine est par ailleurs prolifique en terme de modèles et de vocabulaires, avec par exemple MARC, CIDOC/CRM, FRBRoo, FRBR-ER, VIAF, DublinCore, RDA, FaBIO, BIBFRAME, LD4L, EDM (Europeana). S'ajoutent également des ontologies plus spécifiques comme MusicOntology et MusicBrainz pour le monde musical, BIBO pour le domaine bibliographique ou FOAF pour les relations entre individus.
  Le format MARC (représentation sous forme de notices) est le plus populaire dans le monde bibliographique. Mais il présente quelques inconvénients majeurs. Tout d'abord, des alternatives co-existent (e.g., UNIMARC, MARC21) et limitent ainsi l'échange et la réutilisation directe de notices. De plus, de nombreux champs secondaires (e.g., notes) sont utilisés pour saisir des informations importantes, mais ils ne respectent pas de convention (saisie libre). Chaque institution possède ses propres pratiques de catalogage, et la signification d'une donnée peut donc varier d'une institution à l'autre. Enfin, une notice MARC est une description d'un objet physique, et qui contient donc des informations à différents niveaux : par exemple, le nombre de pages concerne bien l'objet physique, mais le titre de l'oeuvre ou le nom de l'auteur.e se rapportent au travail intellectuel (identique quelque soit l'édition). La communauté a également identifié des motifs bibliographiques récurrents (e.g., traduction, adaptation d'un roman en film), que les notices MARC ne peuvent facilement représenter. En résumé, le format MARC est source de redondance, d'incohérence, d'un manque de granularité et d'une grande complexité, en particulier pour les non-bibliothécaires, et il est donc proscrit pour notre base de connaissances. Le modèle FRBR (Functional Requirements for Bibliographic Records) est plus adapté car il a été validé par le consortium bibliographique comme modèle conceptuel pour une transition sémantique. Cependant, ses diverses implémentations et variantes nécessitent d'étudier leurs différences afin de sélectionner la ou les ontologies pertinentes et qui recouvrent un maximum des informations à représenter.
  De plus, le méta-modèle de notre base de connaissances thématique n'a pas pour objectif de créer une nouvelle ontologie, mais de réutiliser les concepts des ontologies existantes. Pour y parvenir, il est d'abord nécessaire de détecter les concepts et propriétés similaires entre les ontologies existantes. Plusieurs outils d'alignement d'ontologies ont été testés (AgreementMaker, SLog, SMatch et NeON). Ils ont permis de détecter une centaine de correspondances entre les ontologies existantes. Une expertise manuelle montre que de nombreuses correspondantes n'ont pas été découvertes. Ce résultat peu satisfaisant s'explique par le fait que les ontologies existantes ne sont pas au même niveau d'abstraction (certaines étant très générales alors que d'autres sont très spécifiques), et que les outils ont des difficultés pour détecter des correspondances entre une classe et une propriété. Parfois, c'est la relation associée à la correspondance qui est erronée : les outils proposent fréquemment (et souvent exclusivement) une relation d'équivalence, alors que la réalité est bien plus complexe et inclut des relations comme la subsomption, l'inverse, la méréologie, ou la disjonction. Actuellement, nous avons sélectionné six ontologies de base, qui représentent un total de 250 classes et plus de 2000 propriétés. Le méta-modèle comprend actuellement 80 classes et 400 propriétés, et plusieurs centaines de mappings vers les six ontologies. Il permet de représenter la majorité des données concernant l'héritage culturel tout en autorisant un enrichissement (e.g., des liens entre personnes, une hiérarchie de termes descriptifs).
  En 2016 et 2017, nous avons étendu le méta-modèle et vérifié qu'il pouvait être peuplé correctement, à la fois en terme de stockage de données existantes (celles des institutions

culturelles) mais aussi en terme de stockage des données liées à l'enrichissement. Des données existantes (collection de la bibliothèque municipale de Vaulx-en-Velin, catalogue des Hospices Civils de Lyon) ainsi que de nouvelles données (sources externes du Linked Open Data, comme DBpedia, Wikidata et DataBNF) ont été utilisées pour construire des premières bases de connaissances et valider ce méta-modèle. Les expérimentations sur données existantes sont amplement concluantes. Celles sur les données d'enrichissement ont été réalisées sur de petits jeu de données et nécessitent donc plus d'approfondissement.
Un aperçu de ce méta-modèle est montré sur la figure 1 (ci-dessous). Il est disponible en format graphique et en format exploitable par un programme informatique à cette URL : http://research.progilone.fr/mediawiki/index.php?title=Home
Personnes impliquées : Audun Vennesland, Trond Aalberg, Joffrey Decourselle, Fabien Duchateau et Nicolas Lumineau.

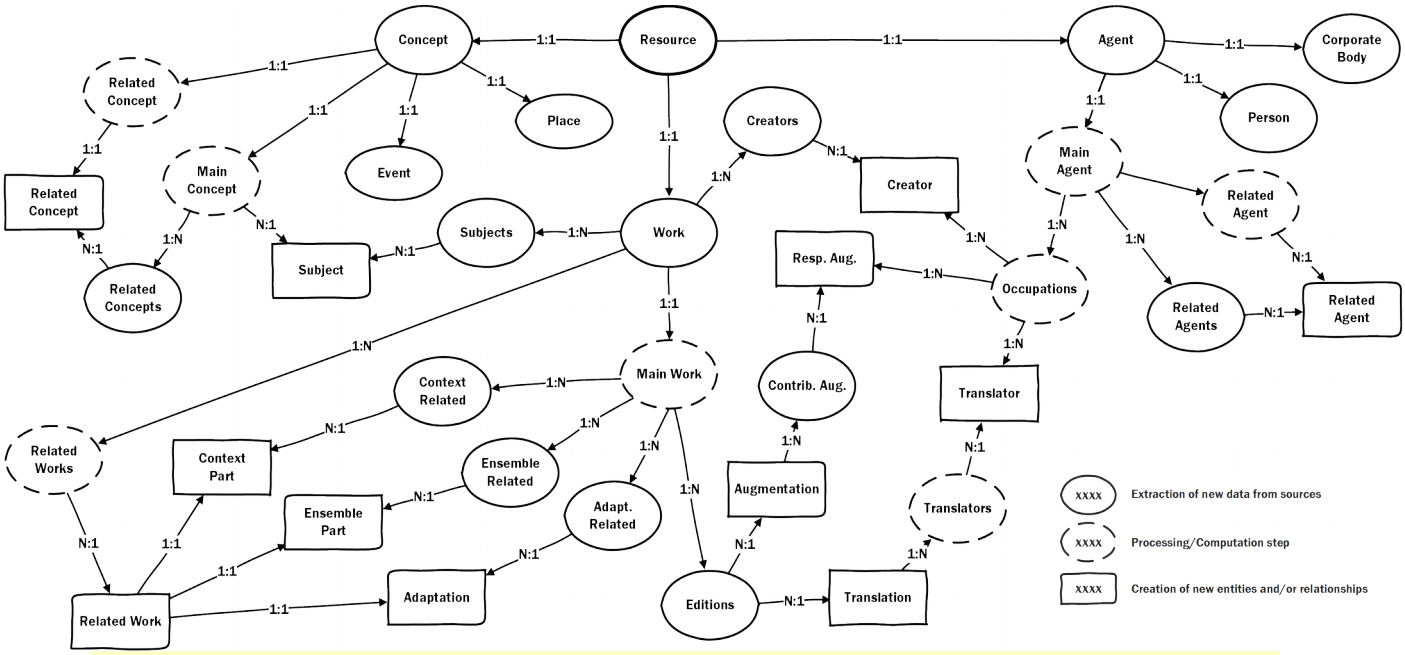

Figure 1. Extrait du méta-modèle utilisé par la base de connaissances et par le processus de FRBRisation. On retrouve notamment les concepts principaux du modèle FRBR (Work, Agent, etc.) ainsi que les motifs bibliographiques principaux (RelatedWork, Translation, etc.)

- En 2015, nous avons également démarré la construction du benchmark. Ce travail s'est prolongé pendant les trois années du projet PICS. Etant donné le contexte (héritage culturel) et les contraintes liées à la thèse CIFRE de Joffrey Decourselle, le jeu de données se devait d'inclure le format le plus populaire dans les institutions culturelles, à savoir MARC (MAchine Readable Cataloguing). Bien que structuré, le format MARC n'est pas suffisant pour représenter toute la sémantique liée aux données culturelles. Il a donc été décidé d'utiliser en complément le modèle FRBR.
  La contribution principale est le benchmark BIB-R, pour « Benchmark for the Interpretation of Bibliographic Records ». Il regroupe des jeux de données en MARC et FRBR et 38 métriques d'évaluation de la FRBRisation. Il est donc utile pour la communauté recherche en informatique et bibliothèques numériques, car il n'existe actuellement aucun jeu de données standard pour tester le processus de FRBRisation et ses différentes implémentations. De plus, ces jeux de données en MARC et FRBR servent de sources de données pour la génération de bases de connaissances thématiques. L'un des jeux de données (T42) permet d'évaluer un outil de FRBRisation pour chaque motif bibliographique ou pour chaque caractéristique du domaine. Le second jeu de données (BIB-RCAT) porte sur un catalogue réel, dans lequel les notices contiennent potentiellement différentes caractéristiques et motifs. Ces jeux de données permettent aussi de mesurer la capacité des outils à intégrer des informations de source

externes, notamment en proposant des notices dont les informations ne sont pas suffisantes pour produire une FRBRisation correcte. Des expérimentations ont été menées avec trois outils afin de comparer leurs performances et la qualité résultant de la FRBRisation des différents jeux de données. Le benchmark comporte 38 métriques d'évaluation qui se situent à trois niveaux : en amont du processus (préparation des données, configuration de l'outil), pendant le processus (interprétation et déduplication), et après la FRBRisation (évaluation de la qualité du catalogue sémantique produit). Les métriques en amont donnent des indices pour configurer l'outil, et notamment les règles de migration entre MARC et FRBR. Elles permettent également de vérifier que chaque champ important sera interprété par le processus. Les métriques qui évaluent la FRBRisation s'intéressent par exemple au temps d'exécution, qui est un aspect crucial lorsque les catalogues sont volumineux. Enfin les métriques post-FRBRisation évaluent le catalogue FRBRisé en le comparant à un « gold standard » en termes de motifs détectés, de sémantique différentes, d'informations erronées ou manquantes. L'ensemble de ces métriques permet de juger des capacités et de comparer des outils de FRBRisation.

Le benchmark a été valorisé en 2016 à travers deux publications (JCDL 2016 et TPDL 2016) puis par un article étendu dans la revue IJDL (accepté en novembre 2017, parution en 2018). Cette dernière version ajoute de nouvelles expérimentations et de nouvelles métriques d'évaluation.

Le benchmark est disponible pour la communauté à cette adresse : http://bib-r.github.io/

Personnes impliquées : Joffrey Decourselle, Fabien Duchateau, Trond Aalberg, Naimdjon Takhirov (Westerdals University, Oslo) et Nicolas Lumineau.

- La dernière tâche ("entity extraction") consiste à définir des "parsers" pour extraire les entités des différents types de sources de données. Cela nécessite l'utilisation de l'alignement d'ontologies décrit précédemment, mais également un parser capable d'extraire correctement les informations du format le plus populaire du monde bibliographique, MARC. La conversion de données MARC en FRBR (FRBRisation) est un processus ambitieux mais nécessaire pour exploiter la multitude de collections présentes dans les bibliothèques ou les musées. Un outil de FRBRisation à destination des institutions culturelles facilite la transition vers un modèle sémantique. La transformation d'une notice MARC est une tâche difficile à cause de l'interprétation de certains champs et de leur valeur. En effet, des informations importantes peuvent être manquantes ou des pratiques locales de catalogage peuvent être utilisées. Par exemple, pour une bande-dessinée, la scénariste apparait avec la responsabilité principale en tant qu'auteure, mais la dessinatrice, qui est stockée comme responsable secondaire, ne possède pas de sous-champ pour identifier son rôle. De la même manière, la catégorie d'un livre peut contenir la valeur « r », un code interne à la bibliothèque attribué aux romans. Dans de telles situations, il est difficile de transformer ces données correctement. Une autre difficulté pendant l'interprétation des notices MARC concerne la détection des motifs bibliographiques (que l'on retrouve également pour les autres institutions culturelles). Le motif le plus simple se compose d'une seule oeuvre, d'une seule expression (édition), d'une seule manifestation (représentation physique) et d'une personne auteure de l'oeuvre. En complément, on distingue quatre catégories de motifs bibliographiques plus complexes : les *augmentations* spécifient un contenu additionnel à une oeuvre, mais dont le degré d'importance est faible (e.g., préface ou illustration de couverture) ; les *dérivations* sont des oeuvres issues d'une modification d'une autre oeuvre (e.g., traduction, adaptation) ; les *agrégations* représentent une relation d'ensemble avec ses composants (e.g., le seigneur des anneaux, ou des pistes musicales dans une compilation) ; les *oeuvres complémentaires* modélisent des oeuvres liées et de même importance (e.g., parodie, manuel d'exercice et son corrigé). La détection de ces motifs nécessite généralement d'interpréter correctement plusieurs champs. Il est donc fréquent de ne découvrir qu'une partie du motif, ce qui empêche ou restreint la FRBRisation du motif. Pour résoudre ces problèmes, nous avons proposé de considérer la FRBRisation à un niveau d'abstraction supérieur, c'est à dire qu'au lieu de ne prendre en compte que le modèle d'entrée (MARC) et le modèle de sortie (FRBR), nous

utilisons un méta-modèle qui permet de mettre en correspondance les informations du modèle en entrée avec celles attendues dans le modèle de sortie (intégration de données). Ce méta-modèle permet de simplifier la FRBRisation en limitant la redondance et en garantissant une interprétation plus correcte des motifs les plus complexes. Il sert également d'artefact concret pour communiquer entre chercheurs en informatique et en bibliothèques numériques. Ci-dessous nous présentons un extrait de ce méta-modèle. Une version implémentée, au format XML, permet de l'intégrer dans un outil de FRBRisation.

En complément de la conversion de notices MARC en FRBR, nous exploitons d'autres sources de données externes. Celles-ci ne ne conforment pas au même vocabulaire ni aux même contraintes. Nous avons décidé d'aligner chaque source de donnée avec notre méta-modèle (qui sert ainsi de pivot), plutôt que d'aligner directement deux sources de données. Cela permet de stocker directement les entités extraites dans notre base de connaissances tout en s'assurant que les alignements directs entre sources de données soient obtenus par transitivité. Actuellement, l'extraction ne se focalise que sur trois sources de données externes : DBpedia, Wikidata et DataBNF. Cette restriction nous permet de détecter les éventuels problèmes et de construire le flot complet du processus pour parvenir à non seulement extraire une entité, mais aussi l'aligner correctement (déduplication) et enfin l'intégrer à la base de connaissances existante. La fusion de données est un problème complexe (non prévu par le projet). Cependant, nous avons tout de même développé un module permettant la fusion de valeurs proches selon différents indices, notamment pour ne pas détériorer la qualité de la base thématique (multiples redondances d'une même information). Dans notre contexte, il est également important que l'utilisateur / utilisatrice décide des informations à visualiser, à explorer et à valider.

Dans les résultats préliminaires, nous avons constaté l'intérêt de construire une base de connaissance thématique : pour enrichir une oeuvre littéraire contenue dans un catalogue en FRBR, une première source (DBpedia) fournit de nouvelles informations sur les liens avec d'autres oeuvres (e.g., une suite, un prélude, le métier de son créateur/créatrice) tandis qu'une autre source (DataBNF) sera capable de donner des informations plus précises sur l'oeuvre initiale en tant que concept littéraire (e.g., des traductions). Ces premiers résultats montrent déjà que l'extraction d'entités pourrait être orientée en priorité vers certaines sources selon le contexte, la requête ou le profil utilisateur / utilisatrice. Un stagiaire (Grégory Howard) a été recruté pendant l'été 2016 pour développer ces parsers sous la supervision de Joffrey Decourselle. Des réunions hebdomadaires étaient planifiées avec Nicolas Lumineau et/ou Fabien Duchateau pour suivre l'avancement du stage.

L'outil de FRBRisation a été valorisé par une affiche présentée en 2016 (poster TPDL). Il est inclus comme module dans le logiciel industriel Syrtis. Une version de démonstration est disponible à cette URL : http://demo-research.progilone.fr/home

Le méta-modèle est disponible en format graphique et en format exploitable par un programme à cette URL : http://research.progilone.fr/mediawiki/index.php?title=Home

Personnes impliquées : Joffrey Decourselle, Grégory Howard, Fabien Duchateau et Nicolas Lumineau.

Pour le work package WP2 (matching) :

Ce work package se découpe en 5 tâches qui concernent l'alignement d'ontologies (aspect qualitatif et aspect passage à l'échelle), l'alignement d'entités (aspect qualitatif et aspect passage à l'échelle), et enfin la partie évaluation de ces approches.

- L'alignement d'ontologies est utilisé à deux niveaux dans le projet. Tout d'abord, c'est un processus qui facilite la construction de l'ontologie (méta-modèle) de nos bases de connaissances. Comme le méta-modèle doit être correct et vérifié manuellement, les outils existants ont été utilisés afin de détecter des correspondances initiales, et ensuite l'alignement a été corrigé et complété (voir la partie WP1 - méta-modèle).
  Dans un deuxième temps, l'alignement est utilisé lorsqu'une base de connaissances est construite. Il n'est pas possible (et pas souhaitable) d'inclure dans notre ontologie globale tous les concepts de toutes les sources de données possibles. Aussi il faudra détecter à la volée certaines correspondances entre notre méta-modèle et une source de données (voir WP1 - entity extraction).
  Sur le plan qualitatif, Audun Vennesland travaille sur l'infrastructure COMPOSE pour l'alignement d'ontologies plutôt que la réutilisation d'un outil existant. En effet, l'un des problèmes avec les approches existantes concerne l'absence de prise en compte des caractéristiques propres aux ontologies (alignement des propriétés et détection de relations autres que l'équivalence). Dans notre contexte, il est important de détecter correctement la relation entre deux concepts pour générer la base de connaissances, car cette relation a un impact sur la fusion de données et l'interrogation. L'approche proposée repose sur une analyse des ontologies en entrée afin de détecter les mesures de similarité pertinentes à appliquer, y compris celles détectant la subsomption. Une fois ces mesures identifiées, l'approche COMPOSE les combine selon différentes stratégies (séquentielle, séquentielle pondérée, parallèle ou hybride). Ce travail permet d'améliorer la qualité de l'alignement, notamment en terme de relation impliquée entre deux concepts. De plus, la sélection automatique de mesures pertinentes est un atout pour gagner en performance, puisque les mesures de similarité inutiles (pour deux ontologies fournies en entrée) ne seront pas appliquées.
  Audun Vennesland est auteur d'une publication sur la détection de relations autres que l'équivalence avec l'approche COMPOSE (conférence WI, 2017).
  Personnes impliquées : Audun Vennesland, Trond Aalberg.

- Concernant l'alignement d'entités, une première approche avait été proposée en 2015 dans un contexte de déduplication, i.e., la détection d'entités similaires dans une même source de données. L'avantage de commencer par une déduplication plutôt que par un alignement d'entités entre plusieurs sources est de connaître les classes/propriétés qui sont comparables. Ainsi, nous nous focalisons complètement sur le processus d'alignement d'entités sans dépendre des résultats d'un autre processus (alignement d'ontologies).
  Pour les aspects performances, nous avons adapté un algorithme distribué existant (Pair Range) afin d'équilibrer le travail d'alignement de chaque machine. La différence principale avec Pair Range réside dans le fait que cette approche considère toute comparaison de deux entités comme équivalente (en temps de traitement) lors de la répartition des comparaisons sur les différentes machines. L'adaptation consiste donc à différencier les comparaisons (nombre de propriétés à comparer, taille des valeurs, etc.), puis de les répartir équitablement sur les machines de traitement selon un algorithme de « bin packing ».
  Pour les aspects qualitatif, la détection d'entités similaires s'effectue traditionnellement en comparant un sous-ensemble d'attributs. La problématique principale est la sélection de ce sous-ensemble. Contrairement aux approches existantes, nous pensons qu'il est envisageable de sélectionner un sous-ensemble différent pour chaque paire d'entités à comparer. Cette contribution a d'abord était testée en mode semi-automatique (une pré-sélection d'attributs est réalisée), puis d'autres améliorations ont été apportées. Tout d'abord, la prise en compte du contexte d'intégration (en plus de la déduplication), i.e., le processus aligne des entités issues

de différentes sources. La qualité de cet alignement dépend donc de l'alignement des modèles / ontologies. Comme expliqué dans la partie WP1 - "entity extraction", notre méta-modèle sert de pivot pour détecter les correspondances entre concepts des différentes sources. Des expérimentations sont toujours en cours, notamment pour mesurer l'impact de l'alignement ontologique sur la qualité de l'alignement d'entités. Un second point concerne les itérations pendant l'algorithme d'alignement. Elles permettent désormais de continuer l'alignement aux concepts suivants lorsque des concepts ont été alignés. Par exemple, l'alignement de deux oeuvres va déclencher, en exploitant notre méta-modèle, des tentatives pour détecter des correspondances entre les propriétés de ces oeuvres, mais aussi entre les autres entités reliées à ces oeuvres (e.g., la personne créatrice de l'oeuvre). L'algorithme s'arrête quand toutes les branches actives ont été explorées. Enfin, l'automatisation de la partie qualitative permet de sélectionner automatiquement la meilleure clé de blocking pour les paires d'attributs. Ces derniers sont regroupés dans un treillis et pour une paire d'entités donnée, l'algorithme sélectionne automatiquement le sous-ensemble d'attributs le plus intéressant pour comparer cette paire. L'intuition est que ce sous-ensemble permet une répartition plus fine des paires d'entités à comparer, et chaque paire est également comparée avec le sous-ensemble d'attributs communs maximal. L'impact sur les performances doit encore être évalué car le pré-traitement (calcul des statistiques sur les paires, et sélection de la clé de blocking) a un impact négatif sur les performances, qui - dans l'idéal - doit être compensé par le traitement (meilleure répartition et moins de comparaisons).
Un article de 8 pages a été rédigé sur l'adaptation de Pair Range et la sélection d'une meilleure clé de blocking. Cependant, la section expérimentations n'est pas encore terminée et ces travaux seront repris en 2018 (et idéalement soumis à une conférence).
Personnes impliquées : Joffrey Decourselle, Fabien Duchateau et Nicolas Lumineau.

- Une autre piste de recherche a émergé pendant ce projet, et nous avons commencé à l'explorer. De nombreux outils d'alignement existent (pour les ontologies ou les entités). Puisque ces outils sont intrinsèquement différents, ils génèrent des alignements différents (contenant des correspondances correctes, les vrais positifs, mais également des incorrectes, les faux positifs, et des manquantes, les faux négatifs). L'idée consiste à se demander s'il serait pertinent de combiner les alignements de différents outils afin d'en améliorer la qualité. En particulier, est-ce que la combinaison permet de défausser des faux positifs ou au contraire de promouvoir des faux négatifs, voire de faciliter la détection de correspondances complexes ? Serait-il envisageable de définir des règles ou motifs qui permettent d'améliorer la qualité lorsque l'on dispose de plusieurs alignements ? Pour répondre à ces questions, nous nous sommes d'abord intéressés à l'alignement d'ontologies, et notamment la campagne OAEI qui met en compétition plusieurs outils sur différentes jeux de données. En 2016, les alignements détectés par une quinzaine d'outils pour le jeu de données « benchmark » (contenant une centaine de tests, i.e., paires d'ontologies à aligner) ont été mis à disposition. Nous avons donc testé différentes stratégies de combinaison (union, intersection, intersection par niveau, intersection pondérée, simple vote en exploitant la relation, largest subgraph, etc.). Les résultats préliminaires sont intéressants : notre approche se classe 3ème par rapport aux quatorze outils (en terme de F-score). Nous constatons également qu'il ne faut pas énormément d'outils pour obtenir ce résultat (4 ou 5 suffisent), et que nos correspondances proviennent effectivement de la combinaison de différents alignements. Cependant les deux outils qui obtiennent un meilleur score sont fortement configurés pour ce jeu de données, et il est difficile de les battre. Nous avons également étudié ce que donnait la combinaison pour l'alignement d'entités. Nous disposons des quatre jeux de données de Rahm et al. (domaine e-commerce et publications scientifiques), soit environ 10 000 correspondances au total pour des collections contenant entre 1100 et 65000 entités. Pour l'alignement d'entités, il est plus difficile de trouver des outils fonctionnels et libres. Mais l'outil FEBRL est générique et permet de simuler différents algorithmes d'alignements. Nous avons donc réutilisé ou défini des algorithmes existants, et générés différents alignements. L'analyse de ces résultats est en cours. Une problématique sous-jacente a concerné la définition de bons algorithmes (en terme

de qualité) pour un jeu de données, étant donné un ensemble de mesures de similarité important (70 disponibles dans FEBRL) et une infinité de combinaisons à travers différentes opérations et seuils. Dans la littérature, nous avons constaté que les études sur les mesures de similarité conseillent effectivement des mesures en fonction des caractéristiques des attributs, mais ces calculs statistiques ne prennent en compte que le rappel (i.e., nombre d'entités détectées par rapport à l'attendu). Or, cela favorise grandement certaines mesures (qui découpent par exemple une chaîne de caractères en petits fragments), et ne reflètent pas la qualité globale puisque la précision n'est pas prise en compte. Une perspective sera donc de proposer une étude plus fine des relations entre mesures de similarité et attributs, en prenant en compte précision et rappel. Une difficulté supplémentaire est l'ancienneté de FEBRL (2006), codé en Python 2, et dont les algorithmes de blocking ne sont pas optimaux.

Un article de 15 pages (simple colonne) a été rédigé sur ces travaux de combinaison pour l'alignement d'ontologies. Comme les résultats nécessitent d'être confirmés avec l'alignement d'entités, l'article n'a pas encore été soumis, et sera retravaillé ces prochains mois pour intégrer les résultats des dernières expérimentations. Nous pensons le soumettre à une conférence au printemps 2018.

Personnes impliquées : Fabien Duchateau et Audun Vennesland.

Pour le work package WP3 (gardening) :

Ce work package consiste en la gestion des bases de connaissances produites. Trois tâches étaient planifiées, concernant le stockage (« knowledge base storage »), la cohérence de la base (« knowledge base consistency ») et la réorganisation des bases de connaissance (« knowledge base organization »). Par manque de temps et de ressources, ce work package a été mis de côté pour se concentrer sur les autres travaux. Les solutions techniques pour les différentes tâches sont décrites ci-dessous.

- L'objectif de la tâche « knowledge base storage » consiste à utiliser une architecture distribuée pour stocker les bases de connaissances et favoriser le passage à l'échelle. Un index distribué type DHT doit permettre l'accès au contenu des bases de connaissances.
  Les bases de connaissances sont stockées dans un SGBD post-relationnel (OrientDB) qui a la particularité de représenter à la fois des graphes et des documents. Il permet donc de modéliser la complexité des relations dans le domaine de l'héritage culturel. De plus ce SGBD est distribué et assure des tâches de maintenance basique (indexation, cohérence, réplication). Bien que le système d'indexation ne prenne pas en compte les spécificités du modèle FRBR, il est complété par l'outil Elastic Search, qui fournit d'excellentes performances pour l'accès aux données.

- La tâche « knowledge base consistency » a pour but de garantir la cohérence des bases de connaissances. L'ajout de mappings ou de données doit ainsi respecter des contraintes définies par exemple dans une ontologie. Notre méta-modèle couvre déjà certaines contraintes du domaine (e.g., une œuvre ne peut être créée que par un agent).

- Enfin, la tâche « knowledge base organization » doit gérer l'évolution des bases de connaissances en les divisant (base trop volumineuse) ou en les fusionnant (bases trop petite). Bien que nos expérimentations n'aient pas montré le besoin de faire évoluer ces bases de connaissances, la question de savoir si de nouvelles données doivent être intégrées dans une base de connaissances ou utilisées pour en créer une nouvelle persiste. Par exemple, imaginons que la thématique soit un livre donné. Un commentaire de texte portant sur ce livre devrait faire partie de la base de connaissances. Mais qu'en est-il d'une critique portant sur le commentaire, ou d'un commentaire (sur une autre œuvre) par le même auteur ? La délimitation du contenu d'une base de connaissances est une problématique complexe. Une vision large favorise des cas d'utilisation comme l'exploration ou la sérendipité, tandis qu'une base limitée facilite par exemple la vérification du contenu. Dans notre contexte culturel, nous avons défini la notion d'entités primaires (e.g., œuvre, expression, agent, lieu). Une base de connaissances porte sur un thème qui correspond à une entité primaire, et l'extension (dans une direction) s'arrête lorsque l'on arrive sur une entité primaire qui n'est ni une Oeuvre, ni une Expression et ni une Manifestation (un niveau de profondeur configurable permettant de ne pas continuer trop loin pour ces types d'entités).

Personnes impliquées : Joffrey Decourselle, Audun Vennesland, Fabien Duchateau Trond Aalberg et Nicolas Lumineau.

Pour le work package WP4 (query processing) :

Dans ce work package, deux tâches étaient planifiées « knowledge base selection » (en 2016) et le traitement de requêtes agrégatives (2017). Cependant, notre vision sur ce WP a évolué suite à différents échanges (notamment les conférences SW4CH en 2015 et TPDL en 2016), et nous présentons dans la suite cette évolution ainsi que le travail réalisé.

- En accord avec Trond Aalberg, nous avons recentré ce travail sur deux scénarios, qui s'activent quand une base de connaissances a été produite. Le premier scénario répond aux besoins des experts du domaine (e.g., les librairies, les archivistes de musées). Dans ce scénario d'enrichissement, un.e expert souhaite ajouter de nouvelles connaissances à des informations existantes (e.g., de nouvelles traductions à une oeuvre, des données personnelles sur l'auteur.e). Le second scénario répond aux besoins des utilisateurs / utilisatrices finales (e.g., usagers de bibliothèque, visiteurs de musée). Pour ce second scénario, c'est la notion d'exploration qui est mise en avant : l'idée est de permettre la découverte de nouvelles connaissances à partir d'un point d'entrée (requête).
  Par rapport aux tâches définies dans le projet, la différence principale porte sur la direction de la requête. Initialement, nous pensions construire une multitude de bases de connaissances potentiellement interconnectées (et sur différents thèmes). Une requête devait donc se propager de l'utilisateur vers les bases existantes afin de détecter les informations intéressantes à récupérer. Dans la nouvelle vision, nous avons préféré générer une base de connaissances à la demande d'un utilisateur / utilisatrice. C'est désormais la requête qui permet la construction d'une base de connaissances.
  Les défis liés à ce WP ont donc également évolué, notamment vers des aspects ergonomiques et cognitifs ("comment représenter le contenu d'une base de connaissances thématique sans noyer l'utilisateur / utilisatrice sous une masse d'informations ?"). Le défi concernant la sélection des bases de connaissances est toujours d'actualité : les bases de connaissances sont stockées et contiennent donc des résultats pertinents pour de prochaines requêtes. Enfin, les aspects distribués sont toujours utilisés pour évaluer les différentes branches d'exploration de la requête (parallélisation des processus).

- Lors de la génération d'une base de connaissances, la première étape consiste à vérifier que la requête n'existe pas déjà. Comme les bases stockées sont déjà enrichies (e.g., noms alternatifs pour les lieux, surnoms ou pseudonyme pour les agents), la requête ne nécessite pas d'extension pour cette vérification. Un moteur d'indexation (Elastic Search) est utilisé pour retrouver rapidement une entité à partir de ses différents libellés. Le même processus est utilisé pour la sélection des bases de connaissances pertinentes pour une requête donnée. Ainsi, les bases de connaissances restent indépendantes et sont connectées à la volée via ce moteur d'indexation. Des expérimentations sur de petits jeux de données ont montré que la base de connaissances se construisait en un temps raisonnable, et permettait effectivement d'obtenir de nouvelles informations. Le premier scénario (enrichissement par un ou une experte bibliographique) est un processus itératif, c'est à dire que différents motifs d'enrichissement sont activables (e.g., à partir d'un agent, trouver son lieu de naissance ou trouver ses œuvres). L'expert.e choisit le motif et une étape d'alignement d'entités est réalisée avec le LOD (Dbpedia, BNF et MusicBrainz). Le résultat se présente sous la forme d'un graphe que l'experte peut survoler pour visualiser les différentes informations (initiales et enrichies). La figure 2 (ci-dessous) illustre une notice MARC (entrée) et son enrichissement avec les sources externes après FRBRisation.
  Une démonstration (4 pages) de ce premier scénario a été soumise à la conférence ISWC 2017. L'outil est disponible ici (il nécessite une étape de FRBRisation en amont) : http://demo-research.progilone.fr/home
  Personnes impliquées : Joffrey Decourselle, Duchateau Fabien, Nicolas Lumineau, Trond Aalberg.

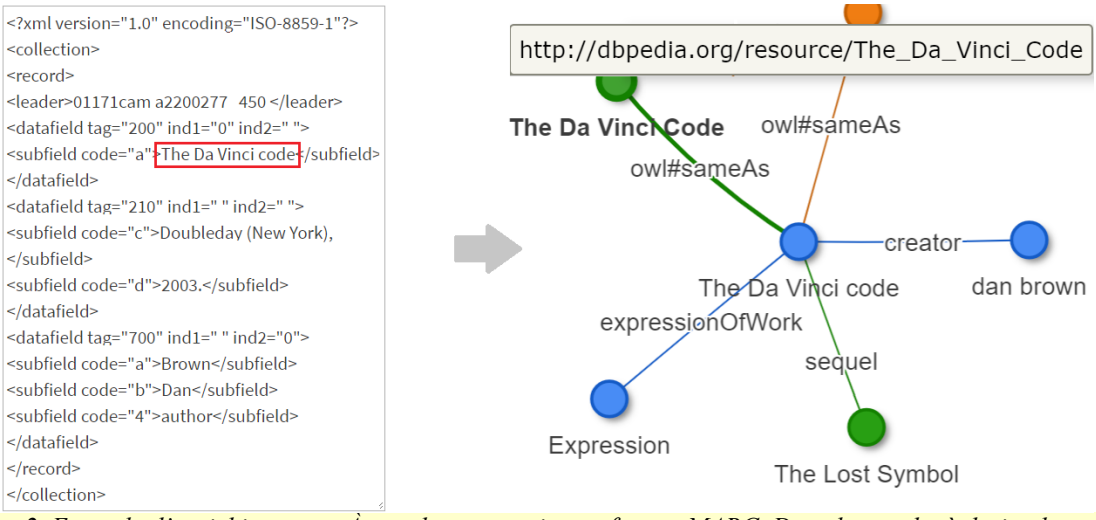

Figure 2. Exemple d'enrichissement. À gauche, une notice au format MARC. Dans le graphe à droite, la version FRBRisée de la notice apparaît sous forme de nœuds bleus. L'enrichissement par les sources externes est montré par les autres nœuds, orange pour BNF et vert pour Dbpedia.

- Pour la partie interfaces, l'objectif est de permettre la navigation et l'exploration dans une collection volumineuse au moyen de filtres et de tris pertinents pour les besoins de l'utilisatrice. Ce scénario s'est focalisé sur des données bibliographiques, puisque le modèle FRBR utilisé pour représenter ce domaine est déjà suffisamment complexe par la richesse et la diversité de ses relations. L'outil est générique et ses filtres sont personnalisés en fonction des collections disponibles. L'idée de calculer la "distance" entre une entité et les termes de la requête n'a pas été retenue, puisque le moteur d'indexation recense les différents libellés pour une entité donnée. Pour l'organisation des différentes informations, les études auprès d'utilisateurs ont montré que ces derniers préféraient utiliser directement le filtres plutôt que de faire confiance à une analyse automatique des termes de la requête (qui nécessite d'être robuste aux variations du langage naturel pour exprimer les informations désirées). Un aperçu de BIBSURF est fourni par la figure 3.
  L'outil BIBSURF est en démonstration ici : http://dijon.idi.ntnu.no/bibsurf/
  Un article présentant l'outil a été publié (JCDL, 2017)
  Personnes impliquées : Trond Aalberg (et des collègues de l'université de Ljubljana, Slovénie).

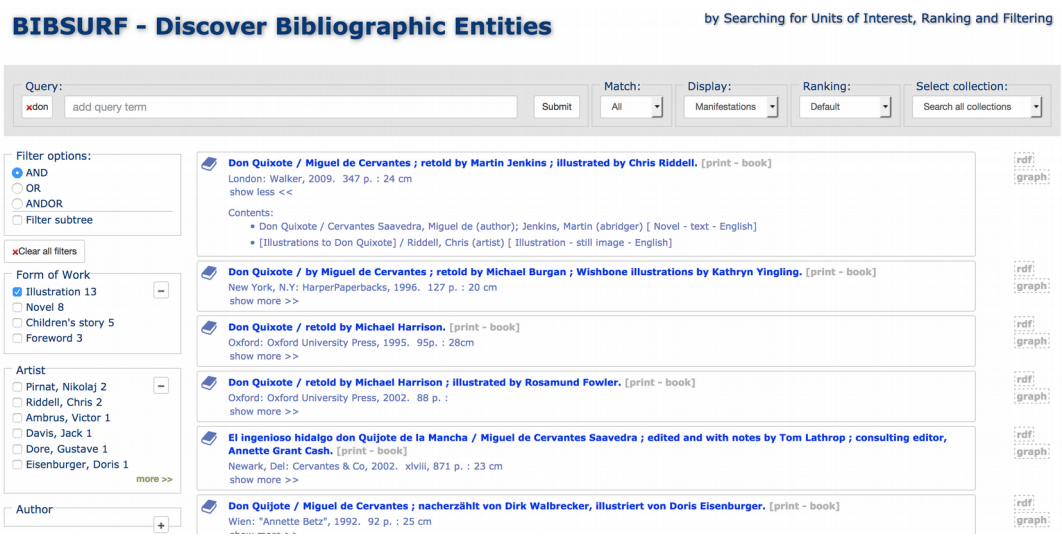

Figure 3. Aperçu de BIBSURF. De nombreux filtres (gauche) permettent d'explorer la base.

Pour le work package WP5 (dissemination) :

Quatre tâches étaient planifiées pour ce work package, à savoir la gestion du site web du projet, la rédaction de publications, la construction d'une plateforme et l'organisation d'un workshop.

- Le site web du projet est maintenu par *Fabien Duchateau* (http://liris.cnrs.fr/diricks/).

- Concernant les publications scientifiques, le consortium a co-écrit 5 publications (dont 1 soumise en attente de notification). Nous avons également publié 6 autres articles sur la thématique du PICS. Deux articles sont en cours de rédaction (un sur l'alignement d'entités, un autre sur la combinaison des alignements), pour lesquels il reste principalement des expérimentations à réaliser et à décrire.
  Personnes impliquées : Joffrey Decourselle, Audun Vennesland, Fabien Duchateau, Nicolas Lumineau et Trond Aalberg.

- La plateforme regroupe actuellement l'outil de FRBRisation, le benchmark pour évaluer la FRBRisation ainsi que le prototype de construction d'une base de connaissances thématiques. Elle est hébergée chez Progilone (entreprise partenaire de la thèse CIFRE de Joffrey Decourselle) pour des raisons pratiques : http://research.progilone.fr/
  Personnes impliquées : Joffrey Decourselle, Audun Vennesland, Fabien Duchateau, Nicolas Lumineau et Trond Aalberg.

- L'organisation d'un workshop n'a pas été possible. L'idée a été discutée lors du séjour de juin 2017, mais les principaux membres du consortium avaient peu de disponibilité pour organiser la gestion d'un workshop à l'automne :
  - Trond Aalberg a obtenu en septembre 2017 un poste de professeur à Oslo (HiOA University), mais il reste également à Trondheim (NTNU) à 25 %. Cela occasionne de nombreux déplacements entre les deux universités (distance de 900 kms).
  - Les enseignant-chercheurs côté LIRIS ont depuis quelques années des charges d'enseignement conséquentes (> 260 heures).
  - Les deux doctorants ont commencé la rédaction de leur mémoire de thèse à partir de septembre 2017 (soutenance printemps 2018 pour Joffrey Decourselle et été 2018 pour Audun Vennesland).

Conclusion et perspectives :

Le projet PICS DIRICKS avait pour ambition de produire et d'exploiter des bases de connaissances thématiques dans le domaine culturel.

Par rapport au projet scientifique initial, deux work packages ont avancé comme prévu (WP1 et WP2). Le WP1 est considéré comme finalisé, tandis que le WP2 nécessite encore quelques expérimentations à plus large échelle. Le work package WP4 a été modifié suite à divers échanges, et les tâches de ce WP4 sont finalisées avec la production d'un outil d'enrichissement et un outil d'exploration. Un work package a été abandonné (WP3), mais des solutions techniques ont été apportées afin de minimiser son impact sur les autres work packages (choix d'un SGBD distribué, moteur d'indexation externe). Cette décision se justifie aussi par la perte d'une ressource humaine sur cette thématique (i.e., le doctorant Kamel Taouche était prévu initialement sur ces tâches, mais il a changé de thématique de thèse avant le début du projet). Enfin, pour le work package dissémination (WP5), la synergie entre les deux équipes aura permis l'acceptation de onze publications (dont cinq co-signées par le consortium) et cinq présentations en conférence (dont quatre co-signées), mais le workshop planifié dans le projet n'a pu être organisé. D'une manière générale, il faut reconnaître que certaines tâches ont pris davantage de temps que prévu : par exemple, l'extraction d'entités (dans WP1) ne prévoyait pas initialement d'exploiter le format MARC, mais sa popularité en a fait un pré-requis pour la construction des bases. De même, la combinaison d'alignements (rattachée et décrite dans WP2) est une piste qu'ils nous a semblé intéressant d'étudier, alors qu'elle n'était pas prévue à l'origine.

Plusieurs travaux issus du projet PICS sont encore valorisables. Il y a d'abord les deux articles en cours de rédaction (un sur l'alignement d'entités, un autre sur la combinaison des alignements). L'article démonstration sur l'enrichissement sémantique pourra être amélioré et soumis à nouveau. La construction du méta-modèle utilisé pour la FRBRisation et pour la cohérence des bases de connaissances est une contribution importante, mais la difficulté pour formaliser ses concepts ont retardé l'écriture d'un article. La rédaction du chapitre de thèse correspondant (pour Joffrey Decourselle) a permis de remettre à plat ces concepts, et il sera envisageable de valoriser les travaux autour de ce méta-modèle.

Le projet PICS offre de nombreuses perspectives sur le plan scientifique et collaboratif.

Au niveau scientifique, la construction du jeu de données dans le benchmark pourrait s'enrichir de sources non structurées (texte). De même les bases de connaissances pourraient, en plus du Linked Open Data, exploiter des sources textuelles, par exemple pour lister des événements en lien avec une œuvre, un lieu ou un agent. L'exploitation des données liées et ouvertes est également utile pour la détection des motifs bibliographiques. Par exemple, DBpedia contient les relations de précédence ou de suite entre plusieurs oeuvres, ce qui permet d'identifier une agrégation. Dans le catalogue FRBRisé de la BNF, les traductions sont fortement présentes et peuvent confirmer un motif de dérivation. Vu l'hétérogénéité des sources externes, la difficulté est d'apprendre à reconnaitre les correspondances nécessaires pour identifier automatiquement les concepts importants d'un motif bibliographique. Le processus de déduplication ou d'alignement d'entités utilise une étape de blocking pour sélectionner un sous-ensemble des paires d'entités qui seront comparées plus finement. Comme le blocking se limite à des algorithmes basiques (e.g., égalité sur l'année de création ou sur le titre de l'oeuvre), des entités équivalentes peuvent être manquées, notamment à cause des nombreux alias (e.g., de personnes, de lieux) ou à cause du multi-linguisme (e.g., pour le titre de l'oeuvre). Les jeux de données ouverts et liés comme VIAF ou DBpedia offrent une liste de noms alternatifs ou des titres dans plusieurs langues, et pourraient donc être utilisés pour améliorer ces processus. L'une des perspectives les plus attendues porte sur une application qui valorise la base de connaissances auprès des utilisateurs, voire autorise leur partage entre différentes institutions. Par exemple, les bibliothèques manquent d'outils pour visualiser le catalogue FRBRisé et les expert.e.s ont besoin de corriger facilement les erreurs pouvant résulter de la migration (outil d'annotation collaboratif). L'apprentissage permettrait de reconnaitre puis de suggérer d'éventuelles erreurs. Enfin, l'utilisation

de la base de connaissances peut être analysée pour savoir comment sont consommées les données. Cette analyse peut porter sur le type de requêtes, les ressources demandées, les chemins suivis lors de l'exploration, la manière d'enrichir, les scénarios utilisateur, etc.

Au niveau collaboration, nous avons mentionné le montage d'un projet européen en 2017 (voir B.3). Les membres « fondateurs » de ce projet (Norvège, Allemagne, Slovénie, France) ont des compétences complémentaires et nous envisageons de répondre à nouveau à un appel européen (DT-TRANSFORMATIONS-12-2018-2020). Des discussions pour cibler les applications potentielles et les éventuels partenaires ont déjà eu lieu en novembre et se poursuivront début 2018.

B.2 - Co-encadrement de doctorants et/ou participation à des jurys

a) Thèses co-encadrées ou en co-tutelle transnationale
Titre de la thèse, nom du doctorant, laboratoire principal de rattachement, nom des co-encadrants dans chaque laboratoire.

Il n'y a pas de co-encadrement officiel pour les deux thèses liées au projet PICS (Joffrey Decourselle pour le LIRIS et Audun Vennesland pour le NTNU).
Cependant, les deux doctorants sont suivis et conseillés par un chercheur du laboratoire partenaire (Trond Aalberg pour Joffrey, et Fabien Duchateau pour Audun). Cet accompagnement se traduit par exemple par quatre publications conjointes.

b) Participation à des jurys de soutenance de thèse ou d'habilitation dans un des laboratoires partenaires étrangers
Titre de la thèse/habilitation, nom du candidat, laboratoire principal de rattachement, date, lieu de la soutenance, nom du (des) membre(s) du PICS participant au jury

B.3 – AUTRES ACTIVITES COMMUNES
Activités avec des chercheurs du laboratoire partenaire étranger hors du contexte du PICS, projets co-déposés dans le cadre d'appels nationaux ou européens, contrats industriels,…
Objet, cadre, dates, bref descriptif.

Un pré-projet intitulé CHECK (Cultural Heritage Extraction to Construct Knowledge) a été déposé en février 2017 lors de l'appel à projet européen « CULT-COOP-09-2017: European cultural heritage, access and analysis for a richer interpretation of the past ». Le consortium rassemblait des laboratoires/universités de Norvège, France, Allemagne, Slovénie, Royaume-Uni, ainsi que deux partenaires industriel et associatifs (Royaume-Uni et Italie).
L'objectif de ce pré-projet est de faciliter les échanges sur les connaissances autour des objets culturels en combinant des sources de données pertinentes et en proposant des outils innovants pour l'enrichissement collaboratif des métadonnées de ces objets par des experts du domaine.
Ce pré-projet n'a passé le premier tour des sélections (13 pré-projets sélectionnés sur 139 déposés). Ce premier appel nous a permis de mieux cerner les attentes pour ce type d'appel.

C. PRODUCTION SCIENTIFIQUE CO-SIGNEE AVEC LES PARTENAIRES ETRANGERS DU PICS

Pour ce dernier rapport, nous rappelons l'intégralité des contributions produites pendant le projet. Elles sont classées par date décroissante, de sorte que celles de l'année 2017 soient listées en premières.

a) Liste des publications parues, acceptées ou soumises (préciser) dans des revues avec comité de lecture

Benchmarking and Evaluating the Interpretation of Bibliographic Records
International Journal on Digital Libraries (IJDL), accepté en 2017, à paraitre en 2018
Trond, Aalberg and Duchateau, Fabien and Takhirov, Naimdjon and Decourselle, Joffrey and Lumineau, Nicolas

Impact des données ouvertes et liées sur les catalogues bibliographiques
Ingénierie des Systèmes d'Informations (ISI), soumis en 2017
Duchateau, Fabien and Lumineau, Nicolas and Trond, Aalberg

b) Liste des publications dans des ouvrages (livres, proceedings, … préciser)

Open Datasets for Evaluating the Interpretation of Bibliographic Records
Joint Conference on Digital Libraries (JCDL, rang A+), parue en 2016
Decourselle, Joffrey and Duchateau, Fabien and Aalberg, Trond and Takhirov, Naimdjon and Lumineau, Nicolas [hal-01302830v2]

BIB-R: A Benchmark for the Interpretation of Bibliographic Records
Theory and Practice of Digital Libraries (TPDL, rang A), parue en 2016
Joffrey Decourselle and Fabien Duchateau and Trond Aalberg and Naimdjon Takhirov and Nicolas Lumineau [hal-01324529]

A Novel Vision for Navigation and Enrichment in Cultural Heritage Collections
New Trends in Databases and Information Systems: ADBIS (2015), p. 488
Joffrey Decourselle, Audun Vennesland, Trond Aalberg, Fabien Duchateau and Nicolas Lumineau [hal-01194308]

c) Liste des présentations à des colloques co-signées avec les partenaires étrangers du PICS
(indiquer si exposés oraux ou affiches)

Open Datasets for Evaluating the Interpretation of Bibliographic Records (affiche)
Joint Conference on Digital Libraries (JCDL, rang A+), 2016
Decourselle, Joffrey and Duchateau, Fabien and Aalberg, Trond and Takhirov, Naimdjon and Lumineau, Nicolas [hal-01302830v2]

BIB-R: A Benchmark for the Interpretation of Bibliographic Records (exposé oral)
Theory and Practice of Digital Libraries (TPDL, rang A), 2016
Joffrey Decourselle and Fabien Duchateau and Trond Aalberg and Naimdjon Takhirov and Nicolas Lumineau [hal-01324529]

Case-oriented Semantic Enrichment of Bibliographic Entities (affiche)
Theory and Practice of Digital Libraries, 2016, Hannover, Germany
Joffrey Decourselle [hal-01346830v2]

A Novel Vision for Navigation and Enrichment in Cultural Heritage Collections
Semantic Web For Cultural Heritage (SW4CH), septembre 2015, Poitiers (France)
Joffrey Decourselle, Audun Vennesland, Trond Aalberg, Fabien Duchateau and Nicolas Lumineau [hal-01194308] (exposé oral).

The path towards bibliographic ontologies and linked data (keynote)
International Conference on Metadata and Semantics Research, 2017
Trond Aalberg

d) Liste des brevets en co-propriété

e) Autres co-productions (bases de données, plateformes, sites web, portails thématiques… préciser)

Un benchmark pour l'évaluation des outils de FRBRisation
Ce benchmark comprend deux jeu de données : un synthétique (T42) composé de 42 tests pour mesurer les points forts et faiblesses des outils de FRBRisation, et un jeu de données "réel" (BIB-RCAT) de 600 notices qui simule un catalogue de bibliothèque (i.e., plusieurs "patterns" bibliographiques). Il inclut également les spécifications de l'ensemble des 38 métriques d'évaluation.

La totalité des résultats d'expérimentations de ce benchmark sur trois outils de FRBRisation est également disponible dans un rapport de 45 pages.
http://bib-r.github.io/

Un outil de conversion d'un catalogue MARC vers FRBR
Le format MARC n'étant pas du tout adapté pour être directement utilisé dans une base de connaissances, une transformation des données MARC vers un autre modèle est nécessaire. Le modèle cible choisi est FRBR, qui se trouve à un haut niveau d'abstraction (conceptuel). Contrairement aux outils de FRBRisation existants, un méta-modèle a été défini (au-dessus de FRBR) afin de modéliser des connaissances complexes comme les motifs bibliographiques. L'enrichissement de données en exploitant le LOD est la dernière étape de cette démonstration, et permet, par un processus itératif, la sélection des données à intégrer dans la base de connaissances produites.
Notre outil de FRBRisation et d'enrichissement est inclus comme module dans un logiciel industriel Syrtis. Une version de démonstration est disponible à cette URL : http://demo-research.progilone.fr/home
Le méta-modèle est disponible en format graphique et en format exploitable par un programme à cette URL : http://research.progilone.fr/mediawiki/index.php?title=Home

Towards a Pattern-based Semantic Enrichment of Bibliographic Entities
IEEE TCDL, 2016, 12 (2)
Joffrey Decourselle [hal-01404651]
Bien qu'individuelle au nom du doctorant, cette publication, qui est une version étendue de l'article soumis au Doctoral Consortium de TPDL 2016, résulte des discussions et de l'encadrement des membres du PICS.

Syrtis: New Perspectives for Semantic Web Adoption
BOBCATSSS, 2016, Lyon, France
Joffrey Decourselle, Fabien Duchateau, Ronald Ganier [hal-01258556]
Cette publication dans une conférence organisée par et pour des doctorants a permis à Joffrey Decourselle de confronter ses travaux à un public de libraires. Elle met également en avant le logiciel Syrtis développée par la société Progilone avec laquelle Joffrey effectue sa thèse CIFRE.

BIBSURF: Discover Bibliographic Entities by Searching for Units of Interest, Ranking and Filtering (démo)
Joint Conference on Digital Libraries (JCDL), 2016
Trond Aalberg, Tanja Merčun, Maja Žumer
Ce travail de Trond Aalberg est un premier pas vers l'exploration des bases de connaissances (WP4). En effet, le prototype BIBSURF permet de parcourir de grandes collections en organisant de manière intuitive les nombreuses informations pour l'utilisateur / utilisatrice.

Matcher composition for identification of subsumption relations in ontology matching
Web Intelligence (WI), 2017 [10.1145/3106426.3106503]
Audun Vennesland

A Pattern-based Framework for Best Practice Implementation of CRM/FRBRoo
ADBIS Workshop - Semantic Web for Cultural Heritage (SW4CH), 2015, Poitiers, France
Trond Aalberg and Audun Vennesland and Maliheh Farrokhnia

A Survey of FRBRization Techniques
Theory and Practice of Digital Libraries (TPDL), 2015, pp.185 [hal-01198487]
Joffrey Decourselle, Fabien Duchateau, Nicolas Lumineau

D. OBSERVATIONS
(p.ex. difficultés rencontrées…)

Le doctorant Joffre Decourselle fait une thèse CIFRE, c'est à dire en partenariat avec une entreprise (Progilone, http://www.progilone.fr/). Bien que son sujet de thèse s'inscrive pleinement dans le cadre du projet PICS, il consacre une part de son temps de production à l'entreprise. Les travaux sur la FRBRisation et l'enrichissement ont permis à l'entreprise d'obtenir de nouveaux contrats industriels. L'outil Syrtis, qui inclut ces deux processus innovants, a été installé à la bibliothèque de Vaux en Velin (collection de 100000 notices) et aux Hospices Civils de Lyon (collections rassemblant plus de 2 millions de notices). Un nouveau contrat vient récemment d'être accepté avec le réseau national Canopé pour migrer et enrichir leurs collections avec l'outil Syrtis. Ces réussites valorisent le travail de recherche effectué dans le cadre de la thèse de Joffrey et dans le cadre de la collaboration PICS.

Le doctorant Audun Vennesland est également impliqué dans un autre projet avec l'agence nationale de recherche Norvégienne (SINTEF). Cet autre projet porte aussi sur l'alignement de modèles / ontologies, mais pour le domaine des transports. Audun Vennesland doit donc jongler entre les deux domaines des projets. Entre août 2016 et juillet 2017, Audun Vennesland était à Valencia, Espagne, dans le cadre d'un échange universitaire, et il a donc consacré du temps pour participer aux travaux avec son équipe d'accueil.

Le rapport final du projet PICS est à rendre en décembre de la dernière année. Si cela est possible, il serait préférable de laisser 1 mois après la fin du projet pour rédiger ce rapport, car des réalisations ou des soumissions dans le cadre du PICS peuvent encore avoir lieu en fin d'année.