Semantic Representation of a Heterogeneous Document Corpus for an Innovative Information Retrieval Model : Application to the Construction Industry

Nathalie Charbel

DOCTORAL SCHOOL OF EXACT SCIENCES AND THEIR APPLICATIONS (ED211)

Semantic Representation of a Heterogeneous Document Corpus for an Innovative Information Retrieval Model: Application to the Construction Industry

---

Nathalie CHARBEL

Advisors:	Dr. Christian SALLABERRY	Univ. Pau & Pays Adour, France
	Dr. Sébastien LABORIE	Univ. Pau & Pays Adour, France
	Prof. Richard CHBEIR	Univ. Pau & Pays Adour, France
Reviewers:	Prof. Max CHEVALIER	Toulouse University, France
	Prof. Sylvie CALABRETTO	University of Lyon, France
Examiners:	Dr. Pieter PAUWELS	Ghent University, Belgium
	Dr. Pierre BOURREAU	Nobatek/INEF4, France
	Mr. Frederic Betbeder	Nobatek/INEF4, France
	Prof. Kokou YETONGNON	University of Bourgogne, France

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science

December 21^{rst}, 2018

Try and leave this world a little better than you found it, and when your turn comes to die, you can die happy in feeling that at any rate, you have not wasted your time but have done your best - Robert Baden Powell

I wholeheartedly dedicate this work to my loving parents Ghada and Nabil, and my sweet siblings Carine and Tony for their unconditional love and support.

I also dedicate this work to my best friend and lover Wissam Bejjani, whose constant encouragement, limitless giving and sacrifice helped me accomplish this research.

Above all, to Almighty God who always give me strength, knowledge, and wisdom in everything I do.

Acknowledgements

My research and dissertation would have been impossible without the presence and support of many people and institutions.

Foremost, I would like to express my sincere gratitude to my advisors Dr. Christian Sallaberry, Dr. Sébastien Laborie and Prof. Richard Chbeir. I am deeply thankful to Dr. Christian Sallaberry for his close supervision, constant presence, generous guidance, patience, comprehension and trust during the thesis. I am profoundly thankful to Dr. Sébastien Laborie for his continuous support, motivation, practical suggestions and constructive comments during the completion of my Doctorate. I would also like to express my greatest gratitude to Prof. Richard Chbeir for his warm encouragement, profound belief in my abilities, helpful contributions, valuable advice, exigence and immense knowledge, which inspired me to widen my research from various perspectives. It is a pleasure to have known them and a privilege to have worked with them.

My sincere thanks also go to Prof. Sylvie Calabretto and Prof. Max Chevalier for the time they dedicated to read my dissertation along with their helpful comments. I am also grateful to the rest of my thesis committee Dr. Pieter Pauwels, Dr. Pierre Bourreau, Mr. Frederic Betbeder and Prof. Kokou YETONGON for their valuable questions, insightful comments, and encouragement.

I am immensely grateful to Nobatek/INEF4 which co-funded my thesis. Without its aid and support, it would not be possible to conduct this research. Thank you Mr. Frederic Betbeder, Dr. Pascale Brassier, Dr. Christophe Cantau and Dr. Pierre Bourreau for helping me promote my research in useful industrial applications. I would like to thank the departments of innovative services and technologies for their help in the data collection and the experimental evaluation processes. Special thanks to Mr. Fabian Bertocchi.

I am particularly grateful for the close assistance given by Dr. Gilbert Tekli during the first year of the thesis. It is with pleasure that I acknowledge his efforts and contributions.

Discussions with Dr. Joe Tekli have been illuminating. I have greatly benefited from his knowledge and feedback.

I thank the IUT Bayonne for permission to use the materials and facilities. Special thanks to Prof. Philippe Aniorte and Dr. Philippe Lopisteguy for their kindness and support.

I appreciate the help of Benoit Coumy-Casteret in the implementation of the prototype.

I would like to acknowledge the moral and emotional support of all my colleagues in Nobatek and LIUPPA laboratory, especially Lara Kallab, Elio Mansour, Karam Bou Chaaya, Fawzi Khattar and Joelle Céméli.

I am thankful to my friends in Lebanon, France and abroad for their limitless love, support and availability in the hard moments. Thank you Clea Abi Khalil, Khouloud Salameh and Eliana Raad.

To my closest friend Lara, thank you for listening, offering me advice, and supporting me through this entire process.

I am deeply indebted to my beloved family who has never left me alone despite the distance. Thank you Mom and Dad for supporting me and giving me wings to fly and follow my dreams. Without your drive and support, I might not be the person I am today. Thank you Sister and Brother for being by my side whenever I need.

I owe my deepest gratitude to my love Wissam Bejjani whom without his love, listening, support, inspiration and advice I would not have lasted.

Lastly and most importantly, I am extremely grateful to Jesus Christ who filled me with strength, patience and perseverance to overcome obstacles and finish this work.

Abstract

The recent advances of Information and Communication Technology (ICT) have resulted in the development of several industries. Adopting semantic technologies has proven several benefits for enabling a better representation of the data and empowering reasoning capabilities over it, especially within an Information Retrieval (IR) application. This has, however, few applications in the industries as there are still unresolved issues, such as the shift from heterogeneous interdependent documents to semantic data models and the representation of the search results while considering relevant contextual information.

In this thesis, we address two main challenges. The first one focuses on the representation of the collective knowledge embedded in a heterogeneous document corpus covering both the domain-specific content of the documents, and other structural aspects such as their metadata, their dependencies (e.g., references), etc. The second one focuses on providing users with innovative search results, from the heterogeneous document corpus, helping them in interpreting the information that is relevant to their inquiries and tracking cross document dependencies.

To cope with these challenges, we first propose a semantic representation of a heterogeneous document corpus that generates a semantic graph covering both the structural and the domain-specific dimensions of the corpus. Then, we introduce a novel data structure for query answers, extracted from this graph, which embeds core information together with structural-based and domain-specific context. In order to provide such query answers, we propose an innovative query processing pipeline, which involves query interpretation, search, ranking, and presentation modules, with a focus on the search and ranking modules.

Our proposal is generic as it can be applicable in different domains. However, in this thesis, it has been experimented in the Architecture, Engineering and Construction (AEC) industry using real-world construction projects.

Résumé

Les avancées récentes des Technologies de l'Information et de la Communication (TIC) ont entraîné des transformations radicales de plusieurs secteurs de l'industrie. L'adoption des technologies du Web Sémantique a démontré plusieurs avantages, surtout dans une application de Recherche d'Information (RI) : une meilleure représentation des données et des capacités de raisonnement sur celles-ci. Cependant, il existe encore peu d'applications industrielles car il reste encore des problèmes non résolus, tels que la représentation de documents hétérogènes interdépendants à travers des modèles de données sémantiques et la représentation des résultats de recherche accompagnés d'informations contextuelles.

Dans cette thèse, nous abordons deux défis principaux. Le premier défi porte sur la représentation de la connaissance relative à un corpus de documents hétérogènes couvrant à la fois le contenu des documents fortement lié à un domaine métier ainsi que d'autres aspects liés à la structure de ces documents tels que leurs métadonnées, les relations inter et intra-documentaires (p. ex., les références entre documents ou parties de documents), etc. Le deuxième défi porte sur la construction des résultats de RI, à partir de ce corpus de documents hétérogènes, aidant les utilisateurs à mieux interpréter les informations pertinentes de leur recherche surtout quand il s'agit d'exploiter les relations inter/intra-documentaires.

Pour faire face à ces défis, nous proposions tout d'abord une représentation sémantique du corpus de documents hétérogènes à travers un modèle de graphe sémantique couvrant à la fois les dimensions structurelle et métier du corpus. Ensuite, nous définissons une nouvelle structure de données pour les résultats de recherche, extraite à partir de ce graphe, qui incorpore les informations pertinentes directes ainsi qu'un contexte structurel et métier. Afin d'exploiter cette nouvelle structure dans un modèle de RI novateur, nous proposons une chaine de traitement automatique de la requête de l'utilisateur, allant du module d'interprétation de requête, aux modules de recherche, de classement et de présentation des résultats. Bien que nous proposons une chaine de traitement complète, nos contributions se focalisent sur les modules de recherche et de classement.

Nous proposons une solution générique qui peut être appliquée dans différents domaines d'applications métiers. Cependant, dans cette thèse, les expérimentations ont été appliquées au domaine du *Bâtiment et Travaux Publics* (BTP), en s'appuyant sur des projets de construction.

Chapitre 1

Introduction

De nos jours, les Systèmes d'Informations (SI) mettent à notre disposition des données non ou semi-structurées. Ces données proviennent de documents hétérogènes c.à.d. ayant des contenus multimédia (p. ex., image, texte, audio, vidéo), de formats différents (p. ex., pdf, txt, mp3, png, etc.) et couvrants des sujets divers. Il arrive souvent que ces documents soient liés les uns aux autres par des liens explicites (p. ex., des références à tout ou partie de documents introduites par des auteurs différents) ou implicites (p. ex., selon les thèmes abordés dans les documents). Dans ce contexte, plusieurs exemples d'application peuvent être identifiés : les SIs exploitant les données sur le web, les données personnelles disponible sur nos ordinateurs (emails, notes, photos, etc.), les données échangées dans des projets industriels (p. ex., dans le domaine de l'industrie manufacturière, du bâtiment, de la médecine, de l'agriculture, etc.).

Dans cette thèse, nous nous focalisons sur les projets industriels multidisciplinaires, avec une application particulière sur les projets de construction qui impliquent un échange de documents hétérogènes interdépendants (p. ex., les Cahiers des Clauses Techniques Particulières ou CCTP, les rapports thermiques, les plans 2D, les photos de chantiers, etc.) encapsulant des données non structurées entre plusieurs acteurs ayant des domaines d'expertises et des intérêts différents (p. ex., architectes, ingénieurs, bureaux d'étude technique, etc.) tout au long du cycle de vie d'un bâtiment. Le domaine du BTP constitue un example pertinent car, malgré la transformation radicale liée à l'explosion des technologies de la TIC (p. ex., la maquette numérique du bâtiment appelé *Building Information Modeling ou BIM*) et du Web Sémantique (p.ex, les ontologies modélisant les données du batiment telles que l'ontologie ifcOWL), ce domaine subit une transition numérique relativement lente par rapport aux autres industries. Cela est dû à plusieurs obstacles, parmi lesquels le manque d'un système de RI qui, à partir d'un corpus de documents hétérogènes liés à un projet, retourne des informations pertinentes spécifiques au domaine métier (p. ex., acoustique, thermiques, etc.) ainsi que des informations liées à la structure des documents (p. ex., niveaux de granularités divers tels que les parties pertinentes des documents, dépendances pertinentes entre documents). Cela concerne plusieurs défis scientifiques : (i) le stockage des données, (ii) la représentation des données encapsulées dans les documents hétérogènes, (iii) l'écriture de la requête pour les utilisateurs non experts, (iv) la représentation des résultats de RI avec des informations contextuelles pertinentes, (v) l'optimisation de la RI, (vi) la sécurité des données. Dans cette thèse, nous abordons deux défis principaux dans le cadre d'une démarche de RI novatrice : la représentation des données (*Défi 1*) et la représentation des résultats de recherche (*Défi 2*).

Pour ce faire, nous proposons une nouvelle plateforme générique, applicable

dans plusieurs domaines, intitulée *FEED2SEARCH* (***F***ram***E***work for hybrid mol***E***cule-base***D*** ***SE***mantic ***SEARCH***) qui a pour but de faciliter la RI sémantique pour les utilisateurs non experts à partir d'un corpus de documents hétérogènes. Plus précisément, les contributions scientifiques de la thèse se présentent dans le cadre de cette plateforme et se déclinent principalement en quatre propositions :

- L'ontologie *LinkedMDR* : une nouvelle ontologie multi-couches pour repésenter la sémantique d'un corpus de documents hétérogènes tout en considérant les deux dimensions du corpus (structurelle et métier) ;
- Le graphe sémantique *Tightly Coupled Semantic Graph* : défini formellement pour représenter la connaissance collective d'un corpus de documents multimédias en se basant sur la sémantique de l'ontologie *LinkedMDR* ;
- Les molécules hybrides *Hybrid Molecules* : définies formellement en tant que sous-graphes du graphe sémantique pour représenter une nouvelle structure pour les résultats de la RI tenant compte des informations pertinentes accompagnées d'informations contextuelles couvrant les dimensions structurelles et métiers du corpus ;
- Un nouveau modèle de recherche et de classement à partir de graphe, fondé sur la proposition d'un algorithme appelé *HM_CSA* et des fonctions de classement adaptées, dont le but est de générer les moléules hybrides à partir du graphe sémantique et de les classer convenablement.

Chapitre 2

Représentation Sémantique d'un Corpus Documentaire Hétérogène

Le chapitre 2 présente les premières contributions de la thèse en se focalisant sur le problème de représentation de la sémantique d'un corpus documentaire hétérogène.

Une étude de l'état de l'art autour des standards et des modèles de données existants, qu'ils soient généraux (p. ex., EXIF, TEI, DC, MPEG-7, COMM, M3O, MediaOnt, Mpeg-7 Rhizomik, OntoText Data model, XCDF Data Model et LINDO Data Model) ou bien dédiés au domaine du BTP (p. ex., IFC, COBie, gbXML, ifcOWL, BOT, PRODUCT, OPM, les modèles de données Newforma et Kroqi), est décrite à la lumière des cinq défis suivants : (i) la représentation de liens inter et intra-documents, (ii) la représentation de métadonnées génériques, structurelles et relatives au contenu des documents, (iii) la représentation de la sémantique des informations documentaires, (iv) la représentation de la multi-modalité des documents et (v) la prise en considération de l'évolutivité des informations et des documents.

A la base de la synthèse de l'état de l'art qui montre une limitation des solutions existantes vis à vis des défis dégagés, nous proposons une approche sémantique pour représenter les dimensions structurelles et métier d'un corpus documentaire hétérogène à travers un graphe sémantique appelé *Tightly Coupled Semantic Graph*

tout en répondant aux cinq défis dégagés. Tout d'abord, nous introduisons l'ontologie *LinkedMDR* qui comprend trois couches : (i) la couche noyau (Core) qui joue le rôle de médiateur entre les différentes couches et qui introduit de nouveaux et riches concepts et relations (non définis dans les standards existants), (ii) la couche intégratrice de méta-données standards (Standardized Metadata) liant des descripteurs de standards existants (comme DC, TEI ou MPEG7) et (iii) la couche spécifique au domaine (Pluggable Domain-Specific) qui s'adapte à tout domaine d'application. *LinkedMDR* fournit l'infrastucture nécessaire pour produire le graphe sémantique. Ce denier est ensuite défini formellement. Enfin, nous présentons un algorithme, appelé *Tight Coupling*, qui décrit le processus de génération de ce graphe.

Chapitre 3

Recherche d'Information dans un Corpus Documentaire Hétérogène

Le chapitre 3 aborde le problème de la RI sur un corpus documentaire hétérogène. Il montre l'intéret d'exploiter l'ontologie *LinkedMDR* et le graphe sémantique *Tightly Coupled Semantic Graph* dans un SI tout en fournissant aux utilisateurs des résultats de recherche couvrant des informations contextuelles à la fois structurelles et spécifiques au domaine métier.

Une étude de l'état de l'art est faite sur les modèles de RI classique (p. ex., Booléens, Vectoriels, Probabilistes) et la RI sémantique (p. ex., les modèles de recherche conceptuelle) et les approches qui les implémentent tout en insistant sur les aspects sémantiques (i) *Fully-Fledged*, (ii) orientées *SPARQL*, (iii) orientées graphes, (iv) orientées molécules RDF. Une étude comparative est faite à la lumière des trois défis suivants concernant les résultats de recherche : (i) fournir des niveaux de granularités pertinents, (ii) prendre en considération les dépendences inter et intra-documentaires, et (iii) représenter les résultats dans une structure pertinente avec de l'information contextuelle comprenant à la fois les dimensions structurelles et spécifiques au domaine métier.

A la base de la synthèse de l'état de l'art qui montre une limitation des solutions existantes vis à vis des défis dégagés, nous proposons une nouvelle structure de données, appelée *Hybrid Molecules*, fondée sur la notion de sous-graphes bien formés et issues du graphe sémantique *Tightly Coupled Semantic Graph*. Les molécules hybrides apportent une information essentielle ainsi que des informations contextuelles utiles sur les documents, y compris les dimensions structurelles et spécifiques au domaine. Nous proposons ainsi un nouveau modèle de RI basé sur la notion de molécule hybride. Bien que nous détaillons la chaine complète de traitement de requêtes avec des algorithmes dédiés à l'interprétation de requêtes, la recherche à base de graphe, le classement et la présentation des résultats aux utilisateurs finaux, les contributions de la thèse se situent au niveau des modules de recherche et de classement. De ce fait, nous proposons *HM_CSA*, un nouvel algorithme de recherche basé sur

des graphes qui génère les molécules hybrides à partir du graphe sémantique, ainsi que des fonctions de pondération qui classent les résultats sous forme de molécules hybrides.

Chapitre 4

Evaluation Expérimentale

Le chapitre 4, présente une évaluation expérimentale des contributions de cette thèse.

Nous avons tout d'abord présenté le prototype qui a implémenté, en Java, les différentes couches de la plateforme *FEED2SEARCH* dans le contexte du domaine du BTP. Ce prototype est formé de deux modules principaux. Le premier, intitulé *LMDR Annotator*, annote automatiquement un corpus documentaire hétérogène donné et génère le graphe sémantique *Tightly Coupled Semantic Graph*. Le second, intitulé *HM Query Processor*, utilise le graphe généré pour fournir une liste classée de molécules hybrides en réponse à une requête écrite en langage naturel par l'utilisateur.

Nous avons ensuite présenté un ensemble d'expérimentations menées sur des projets de construction, fournies par le partenaire Nobatek/INEF4 et évalués par des utilisateurs du domaine. D'une part, nous avons comparé la concision de la représentation d'un corpus documentaire hétérogène basé sur l'ontologie *LinkedMDR* avec des modèles de données alternatifs pour la représentation de documents. Nous avons ensuite évalué l'efficacité du module *LMDR Annotator* dans la génération des annotations automatiques. D'autre part, nous avons évalué l'efficacité du module *HM Query Processor* dans la génération des molécules hybrides résultant de l'algorithm *HM_CSA* et des fonctions de pondérations associées. Les expériences ont montré des résultats prometteurs qui nous motivent à poursuivre la mise en œuvre des prototypes et à l'étude de leur efficacité pour une adoption dans des projets réels.

Chapitre 5

Conclusion

Le chapitre 5 conclut cette étude et présente plusieurs axes de recherche futurs que nous prévoyons d'explorer à court et à long terme.

A court terme, nous travaillerons sur l'interface graphique du prototype pour améliorer surtout la présentation des résultats de recherche. Par exemple, dans l'état actuel du prototype, les molécules hybrides sont visualisées à l'aide d'un outil de visualization de graphe (p. ex., NavigOwl) intégré dans le module *HM Query processor*. Nous pensons qu'il faudrait améliorer l'affichage afin d'aider les utilisateurs non experts en informatique à mieux interpréter les résultats de leur recherche. Cela pourrait être fait en explorant l'état de l'art sur les métodologies de *SERP* (Search Engine Result Page) adoptées par les moteurs de recherche actuels et leur impact sur les utilisateurs finaux.

A long terme, les contributions de la thèse seront intégrées dans deux projets européens qui ont déjà démarrés en 2018 : *BIM4REN* avec le partenaire Nobatek et E2S avec le laboratoire LIUPPA. *BIM4REN*[1] vise à fournir un écosystème numérique facilitant l'intégration d'outils numériques innovants, compatibles avec le modèle BIM, dans le processus de rénovation énergétique des bâtiments. Les contributions de la thèse seront intégrées dans le projet pour aider à la RI dans des corpus de documents textuels sur la CVC (Chauffage, Ventilation et Climatisation). Ceci a pour but de fournir aux maquettes BIM vides, générées à partir d'un outil d'analyse 3D lors de la phase de rénovation du bâtiment, les informations manquantes. Quant au projet *E2S*[2], il vise à fournir un SI générique pouvant être configuré, à la demande, par n'importe quelle organisation afin d'intégrer des services multimédias dédiés à l'indexation, au stockage, à l'enrichissement et au traitement de la sécurité des données de différents domaines, telles que les données relatives à l'énergie et l'environnement. Les contributions de la thèse seront intégrées dans le projet et davantage approfondies dans le contexte du Big Data et de la sécurité des données.

[1] `https://www.ef-l.eu/our-projects/bim4ren/`
[2] `https://e2s-uppa.eu/en/index.html`

Contents

List of Figures

List of Tables

List of Acronyms

AEC	**A**rchitecture, **E**ngineering, and **C**onstruction
API	**A**pplication **P**rogramming **I**nterface
BIM	**B**uilding **I**nformation **M**odeling
CAD	**C**omputer-**A**ided **D**esign
CSA	**C**onstrained **S**pread **A**ctivation
DC	**D**ublin **C**ore
GUI	**G**raphical **U**ser **I**nterface
HVAC	**H**eating, **V**entilation, and **A**ir **C**onditioning
ICT	**I**nformation and **C**ommunication **T**echnology
IFC	**I**ndustry **F**oundation **C**lasses
IR	**I**nformation **R**etrieval
LBD	**L**inked **B**uilding **D**ata
MPEG	**M**oving **P**icture **E**xperts **G**roup
NLP	**N**atural **L**anguage **P**rocessing
OWL	**W**eb **O**ntology **L**anguage
PIM	**P**roject **I**nformation **M**anagement
RDF	**R**esource **D**escription **F**ramework
SERP	**S**earch **E**ngine **R**esults **P**age
SIR	**S**emantic **I**nformation **R**etrieval
SME	**S**mall-to-**M**edium sized **E**nterprise
SPARQL	**S**imple **P**rotocol **A**nd **R**DF **Q**uery **L**anguage
SW	**S**emantic **W**eb
TEI	**T**ext **E**ncoding **I**nitiative
URI	**U**niversal **R**esource **I**dentifier
W3C	**W**orld **W**ide **W**eb **C**onsortium
WWW	**W**orld **W**ide **W**eb
XML	e**X**tensible **M**arkup **L**anguage
XSLT	e**X**tensible **S**tylesheet **L**anguage **T**ransformations

Chapter 1

Introduction

"We are drowning in information but starved for knowledge."
- John Naisbitt

1.1 An Insight into Industry 4.0

Over the past two decades, the Digital Revolution has begun to spread its benefits over the developing world [15]. The emergence of Information and Communication Technology (ICT) has spurred opportunities for increased productivity and efficiency in several industries, such as the medical, automotive, aerospace and construction industries. The use of digital technology has been shown to facilitate collaborative approaches to drive innovation and reduce waste [99].

The fourth industrial revolution, commonly known as Industry 4.0, is designed to prepare the industries to adapt to the era of the so called "Smart Factory" which involves Internet of Things (IOT)[1], Cyber-Physical Systems (CPS)[2], Cloud Computing [3] and Cognitive Computing [4] [60]. Along with the surge in these trends comes a large availability of data. This information revolution created a new stage of the Internet development, the Web 3.0, often referred to as, the *Semantic Web* (SW) or the *Web of Data* [11].

"The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation", says Tim Berners-Lee.

The initiatives towards these major transformations of the World Wide Web (WWW) deeply rooted in day-to-day actions of the industries [75]. **The proper handling of the semantic information has become a must for enabling the full potential of Industry 4.0 and its subsidiary fields.** However, moving from raw and unstructured data to intelligent data models that can be semantically reasoned upon remains a major mile stone that is yet to be crossed.

[1] The network of connected physical devices which allows the collection and exchange of data.

[2] A mechanism which provides combination and coordination between physical and computational elements.

[3] The concept of using a network of remote servers hosted on the Internet to store, manage, and process data.

[4] Technology platforms that are based on artificial intelligence and signal processing.

1.2 The Architecture, Engineering and Construction (AEC) Industry

AEC is the term used to represent three different but related entities in one common industry where architects, engineers and contractors work together for the achievement of a common goal: the efficient accomplishment of a construction project. Although the adoption of the advances of ICT has been slow within the AEC industry in comparison to other industries [53], many government agencies and businesses from the private sector are nowadays imposing strategies to accelerate the digital transition within this industry. Below are some examples:

- The French Minister of Territorial Equality and Housing announced, in December 2014, the launch of the so called *"Plan Transition Numérique dans le Bâtiment"* (PTNB) as part of the construction stimulus plan for the purpose of reaching savings of nearly €1 *trillion* in 2025. The dedicated funds are up to €70 *million* [76];
- The UK Government Construction Strategy (GCS) 2016-2020 [101] sets out the Government's plan to develop its capability as a construction client and act as an exemplary client across the industry. One of its principle objectives is to embed and increase the use of digital technology. The resulting projects are worth £163 *billion*;
- Some of the largest construction companies around Europe have joint forces to put together the European Construction Industry Manifesto for Digitalisation[5] which calls on the European Union (EU) to take strong political action for setting the appropriate regulatory framework on data policy and budgetary focus on digital skills, and Research and Development (R&D).

1.2.1 Building Information Modeling (BIM)

BIM is an intelligent 3D model-based process that gives AEC actors (e.g., architects, engineers, consultants, etc.) the insight and tools to more efficiently plan, design, construct, and manage buildings and infrastructure[6] throughout their whole life-cycle (See Fig. 1.1). The use of BIM empowers the collective approach helping actors of the same construction project in exchanging information [54].

BIM covers more than only building geometry, but also various relationships between objects and information beyond geographic data, such as light analysis, quantities and properties of building components. Thus, it allows the actors to perform calculations, analysis and simulations on 3D object properties, which helps them in the process of decision-making [51]. Recent studies invoke seven dimensions of BIM extending its 3D representation to cover (i) the temporal dimension

[5] *Source: Planning & Building Control Today (pbctoday), BIM News, August 2018*, `https://www.pbctoday.co.uk/news/bim-news/digitalisation-construction/46024/`

[6] `https://www.autodesk.com/solutions/bim`

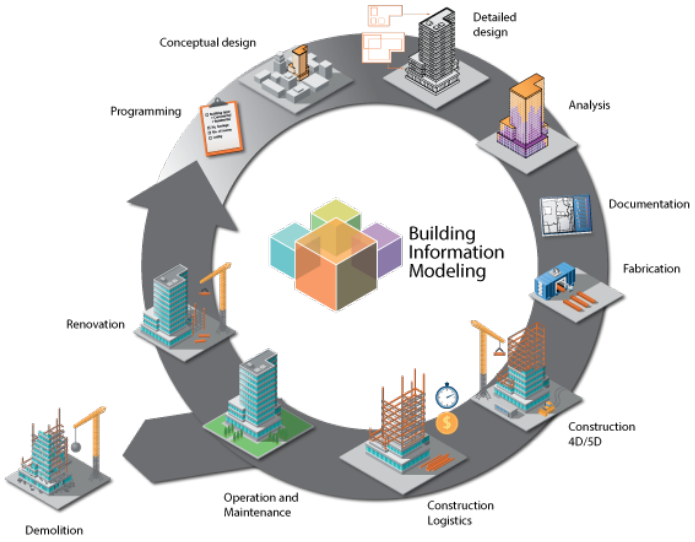


FIGURE 1.1 – BIM throughout the project life-cycle (*Source: The BIM Hub, 2018*).

(4D), (ii) the financial dimension (5D), (iii) the energy dimension (6D), and (iv) the operational dimension (7D) [65].

The first commercial software to implement BIM was ArchiCAD of Graphisoft[7] company. Although many Computer-Aided Design (CAD) vendors implement BIM in their proprietary software solutions, there are initiatives towards a universal approach, such as the openBIM[8] launched by the buildingSMART alliance[9]. The latter aims at ensuring interoperability for the use of BIM across several leading software vendors. Fig. 1.2 shows statistics[10] regarding the rates of BIM adoption in construction projects in some European countries, where France stands in the 3rd position.

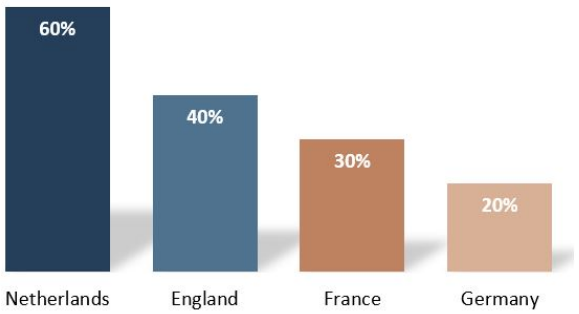


FIGURE 1.2 – Rates of BIM adoption in some European countries.

[7] http://www.graphisoft.com/
[8] https://openbim.fr/openbim/
[9] https://www.buildingsmart.org/
[10] *Source: BIM In Motion, April 2018,* http://biminmotion.fr/ladoption-du-bim-en-europe-la-france-en-troisieme-position/.

1.2.2 Obstacles to Construction 4.0

While BIM presents a major breakthrough in the AEC industry, there are still barriers putting off companies, especially Small-to-Medium sized Enterprises (SMEs), from fully adopting it and keeping up with the recent ICT advances in order to accomplish the shifting to Industry 4.0, also known in the construction industry as Construction 4.0:

- On one hand, the construction industry is a large and old well established sector of the economy making it more resistant to changes relative to other industries. Adopting new technologies, even when the long term advantages are clear, attributes a large upfront cost. Companies are required to invest in acquiring new software tools and professionals. Also, the existing labor force has to be trained to use and rely on new software tools. This problem is aggravated by the lack of enough trained personnel who are qualified to support this transition.
- On the other hand, the uncanny gap between the new technologies and the current status of the industry also plays a major role in hindering the transition to Construction 4.0. For instance, many construction projects rely heavily on the use of documents as they contain interdependent information required to manage the work progress. The BIM model, overlooking the notion of documents, still cannot replace all this information. Further, the users still have to manually intervene in order to enrich the model as there are no existing methods or tools to assist them. As a consequence, the adoption of BIM has somehow been restricted to the early conception phase.

Project Information Management (PIM) solutions [68, 89] are recently being adopted by SMEs as intermediary digitization tools and platforms compatible with the BIM model (by means of plugins, external links to BIM Software). They mainly offer project management services and traditional operations (such as classification, filtering, comparison, and archiving) on documents of several formats. However, they **do not offer any solution that uses the information embedded in the documents to serve the BIM model**.

1.3 Thesis Context

1.3.1 Nobatek/INEF4

Nobatek/INEF4, the co-financier of the thesis, is a French Institute for the Energy Transition of the Building[11]. It provides innovative solutions to support the entire construction sector (architects, contractors, consultants) for the energy and environmental transition.

[11] `https://www.nobatek.inef4.com/`

Since 2014, NOBATEK/INEF4 has been renowned as the French leader for investing in European H2020 projects in the construction sector involving several R&D areas (e.g., Sustainable Districts, Digital Ecosystems, Innovative & accessible BIM tools for all the building renovation value chain). **Nobatek/INEF4 sets out its strategic digital road map for 2024 that emphasizes the development of digital tools and platforms supporting BIM**, and collaborations with active entities in this sector such as laboratories (e.g., Le2i laboratory[12]), enterprises (e.g., EnerBIM[13]) and communities (e.g., Linked Building Data (LBD) Community Group[14] of the World Wide Web Consortium (W3C)).

1.3.2 Motivating Scenario

On the basis of what we presented in previous sections regarding the unresolved semantic information handling issues to enable the full potential of Industry 4.0 in general (See Sect 1.1) and more specifically Construction 4.0 (See Sect. 1.2), we present the following motivating scenario.

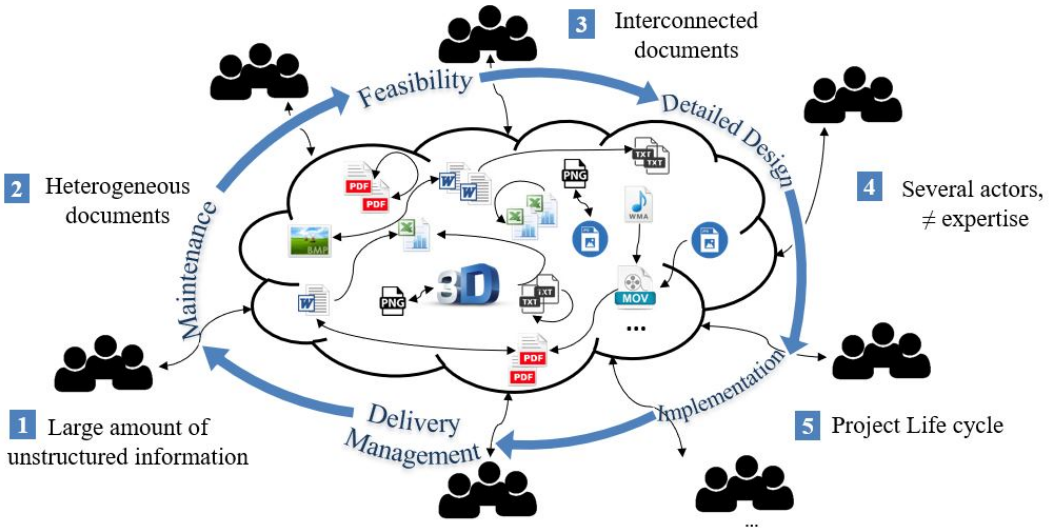


FIGURE 1.3 – Motivating Scenario illustrating a large-scale multi-disciplinary project in the era of Industry 4.0.

Fig. 1.3 illustrates a large scale and multi-disciplinary project involving a large amount of unstructured information dispersed across various heterogeneous documents i.e., documents having different formats, different content, and structure. These documents, although generated from different sources, are interconnected as they have strongly related information (whether related topics, complementary information) that is all together required for the progress of the project. The entities generating these documents are actors of the projects, who have different expertise and intervene in any stage of the project life-cycle.

[12] http://le2i.cnrs.fr/
[13] http://main.enerbim.com/en/partners/
[14] https://www.w3.org/community/lbd/

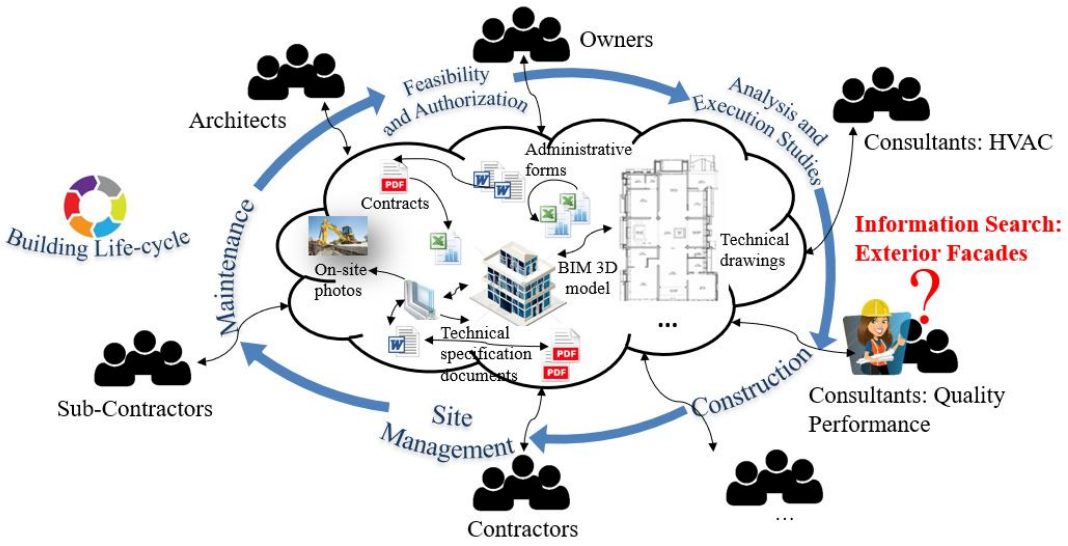


FIGURE 1.4 – Application of the motivating scenario of Fig. 1.3 to the AEC industry in the era of Construction 4.0.

Although the motivating scenario of Fig. 1.3 is applicable to several industries (such as the medical, manufacturing industries), we particularly apply it to the AEC industry (See Fig. 1.4) and give examples on its key elements:

- The multi-disciplinary project comes down to a given construction project;
- The large amount of unstructured information comes down to text and multimedia content describing several topics such as the building data including all the underlying systems (e.g., Heating, Ventilation, and Air Conditioning (HVAC), Electricity, Plumbing, etc.), project costs, etc.;
- The heterogeneous documents come down to contracts, administrative forms, technical specification documents, 2D drawings, BIM files, on-site photos, etc. These documents involve the diverse information described above. They do not follow a common structure and are serialized in different formats (e.g., pdf, docx, xlsx, png, ifc, dwg, etc.);
- The interplay between documents comes down to several implicit and explicit dependencies between them. For instance, a general technical specification document describing general characteristics of the exterior facades of the buildings may refer to a more detailed document describing their thermal properties (e.g., light transmission, solar factor, etc.)
- The actors exchanging these documents come down to owners, architects, contractors, consultants, etc. having different expertise (Architecture, HVAC, etc.)
- The project life-cycle comes down to the building life-cycle starting from the early design stages such as the feasibility (mainly for cost and schedule estimations) and authorization of the project, the analysis and execution studies to the construction, site management and maintenance stages.

In this context, at any point in time of the project, an actor may need to search for a particular information. For instance, consider an engineer who works in a specialized consultancy office supporting the owners of the project in quality performance studies. His main role is to ensure the compliance of the building's exterior facades with environmental standards. For his task to be done, he needs to search for all the information regarding the exterior facades from the whole collection of documents. This means exploring the variety of diverse information related to the exterior facades (such as thermal studies, acoustic studies, etc.) in the form of EXCEL sheets, WORD documents, PDF, technical 2D drawings, etc., finding relevant parts from a bunch of other non relevant information within a single document, and tracking document dependencies to search for complementary or related information.

1.3.3 Main problems

Currently, actors of multi-disciplinary projects have difficulties while searching for a specific information from the collection of heterogeneous documents within these projects. This is due to:

- **Problem 1 -** The lack of an adapted digital tool (e.g., enterprise search engine tool) that is capable of providing the users with the required domain-specific information (e.g., acoustic and thermal studies conducted on the exterior facades of the building) from the heterogeneous document corpus representing the project while considering further structural characteristics of the documents such as the relevant granularity levels (e.g., the section of the document that details on the exterior facades) and the documents' dependencies (e.g., references between the technical specification documents).

  For instance, as mentioned in Sect. 1.2.2, PIM solutions offer traditional operations on documents, such as a keyword-based search over the metadata of documents, which neglects the semantic handling of information contained in these documents and their dependencies;

- **Problem 2 -** The lack of an underlying robust data model which is capable of representing the knowledge embedded in these documents including the domain-specific information and the structural characteristics of the documents.

  For instance, as mentioned in Sect. 1.2.1, although the BIM model is an innovative solution in the AEC industry, it still cannot replace the construction related documents, as it neglects crucial information embedded in these documents (e.g., the dependencies between them).

Consequently, actors of multi-disciplinary projects, especially in SMEs, often manually search for information, which is a very tedious and time-consuming job.

1.3.4 Main Challenges

In this thesis, we mainly address the challenge of searching, in a novel way, for relevant information from the heterogeneous document corpus representing the multidisciplinary project. By novel we mean (i) considering information beyond what is explicitly shown in the raw documents (e.g., their dependencies), (ii) responding to users inquiries by relevant answers helping them to reduce their workload, time and efforts. More concretely, this can result in two main challenges, among several other challenges that remain out of this thesis scope[15]:

- **Challenge 1 -** Representing the collective knowledge embedded in a heterogeneous document corpus through a robust data model;
- **Challenge 2 -** Providing novel search results based on the proposed data model.

Tackling these challenges is crucial and a step forward for reducing the gap between new technologies imposed in the era of Industry 4.0, such as Construction 4.0, and the current status of the market. For instance, in a particular application to the AEC industry, this innovative search could serve the BIM users in enriching the model with the required missing information.

1.4 Proposal

"Semantic Technology makes everything connected to anything and helps to build Linked Data through intricate models for representing information", Source: Ontotext[16]*, 2018*

In this thesis, we first propose to inject semantics on a heterogeneous document corpus and represent it through a semantic network (i.e., knowledge graph using the advances of the *SW*) while considering two different dimensions:

- A structural dimension which represents structural aspects of the documents such as their metadata information, and possible dependencies between them;
- A domain-specific dimension which represents the knowledge describing the content of the documents related to a specific application domain, such as the AEC industry.

We then propose to rely on the resulting knowledge graph, which we call a *tightly coupled semantic graph* for an enriched Information Retrieval (IR) over the heterogeneous document corpus.

1.4.1 FEED2SEARCH Framework

Based on our proposal, we introduce a novel framework, entitled *FEED2SEARCH*, which stands for ***F***ram***E***work for hybrid mol***E***cule-base***D*** ***SE***mantic ***SEARCH*** over a

[15] Such as handling Data Storage, Big Data, Data Security, and Query Writing for non expert users.
[16] `https://ontotext.com/`

heterogeneous document corpus. It is a generic framework since it is applicable to a heterogeneous document corpus in many domains, such as the AEC industry.

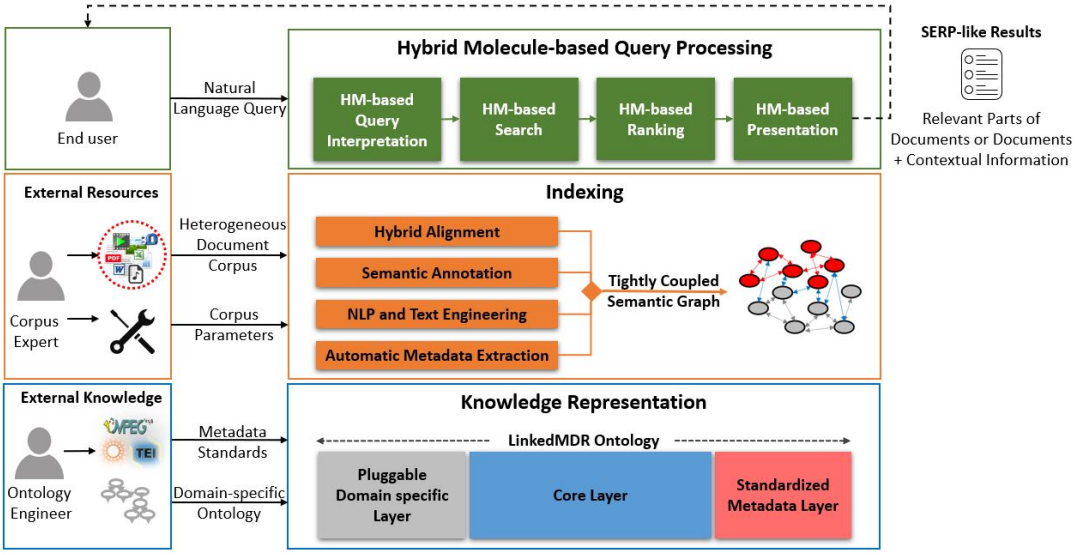


FIGURE 1.5 – An overview of the proposed framework: *FEED2SEARCH*.

The main purpose of *FEED2SEARCH* is to facilitate the query processing over a heterogeneous document corpus for non computer expert users by providing them with relevant answers in response to their natural language queries. As shown in Figure 1.5, *FEED2SEARCH* is made of three interconnected layers where the upper layers use data and services provided by the lower layers:

- **The Knowledge Representation Layer -** It introduces a novel multi-layered ontology, entitled ***LinkedMDR*** [20] which provides the infrastructure for the generation of the *tightly coupled semantic graph*. It is based on external knowledge bases (such as metadata standards and domain-specific ontologies) and handled by an ontology engineer;

- **The Indexing Layer -** It provides services, technologies and Application Programming Interfaces (APIs) (based on well-known techniques such as automatic metadata extraction [42], semantic annotation [26], Natural Language processing (NLP) and Text Engineering [64]) in order to index the given external heterogeneous document corpus and generate the ***tightly coupled semantic graph*** representing it based on the backbone *LinkedMDR* ontology. This layer is handled by a corpus expert which provides the document corpus as well as some parameters required to build the graph (e.g., the granularity levels required to describe the documents);

- **The Hybrid Molecule-based Query Processing Layer -** It provides a comprehensive query processing pipeline, entitled ***Hybrid Molecule-based Query Processing***, based on a novel data structure for query answers, which we call ***Hybrid Molecules***. The latter intervene in each stage of the proposed pipeline and

are constructed progressively from the *tightly coupled semantic graph*, to provide users with relevant information w.r.t. their queries. Consequently given a user's natural language query (expressed as plain text), this layer provides relevant answers in Search Engine Results Page (SERP)-like results on the basis of the information provided by the hybrid molecules.

1.4.2 Main Contributions

The key contributions, previously identified in the different layers of *FEED2SEARCH* framework, are described as follows.

1.4.2.1 *LinkedMDR* ontology

The ontologies provide high semantic expressiveness power for knowledge representation [91]. Thus, *LinkedMDR* ontology is one of the major contributions of this study. The proposed ontology is made of three main layers: the *Core* layer serving as a backbone and a mediator among the different layers, the *Standardized Metadata* layer relying on external metadata standards (e.g., [29, 97, 98]), the *Pluggable Domain-Specific* layer which can adapt to any domain application through a flexible and easy mechanism to align with external domain-specific ontologies, such as well-known ontologies in the AEC industry (e.g. [18]).

1.4.2.2 Tightly Coupled Semantic Graph

The *tightly coupled semantic graph* is one of the major contributions of this study as it represents the collective knowledge embedded in a heterogeneous document corpus based on the infrastructure provided by the proposed *LinkedMDR* ontology. We formally define this graph using two coupled external resources: a heterogeneous document corpus and a domain-specific ontology (for the pluggable domain-specific layer of *LinkedMDR*). Thus, it embeds instances representing the domain-specific components of the corpus, others representing the structural components, and hybrid links representing relations between the two different components. A dedicated *tight coupling algorithm* describes the mechanism of the generation of the graph.

1.4.2.3 Hybrid Molecules

The definition of a novel structure extracted from a tightly coupled semantic graph, which we call the *hybrid Molecules*, is one of the major contributions of this study as it is a means to provide innovative search results covering both the domain-specific and the structural-based dimensions of the documents. The *Hybrid Molecules* consist of well-defined sub-graphs that we formally define in view of the characteristics of a tightly coupled semantic graph and the definition of a molecule concept in the literature [27, 28, 30, 36, 69]. They are hybrid as they encapsulate domain-specific information coupled with related structural-based information of the documents.

The hybrid molecules bring in a core information together with helpful contextual information of the documents.

1.4.2.4 Hybrid Molecule-based Search and Ranking

On the basis of the proposed *Hybrid Molecules*, one of the major contributions of this study is to provide the underlying IR model that exploits this novel structure. Although we present an overall pipeline (*Hybrid Molecule-based Query Processing*) with a bunch of algorithms for each stage of the query processing, we focus in this thesis on the IR model behind the *Hybrid Molecule-based Search and Ranking* modules. The former presents a novel graph-based search algorithm, entitled *HM_CSA* which generates relevant hybrid molecules from a tightly coupled semantic graph w.r.t. a user's natural language query. The latter relies on a series of weight mapping functions which rank the molecules conveniently.

1.5 Publications

The contributions of this thesis are published and submitted to the following conferences:

1. Nathalie Charbel, Christian Sallaberry, Sébastien Laborie, Gilbert Tekli, Richard Chbeir: Un modèle sémantique de représentation de corpus de documents multimédia. INFORSID 2017: 11-26

2. Nathalie Charbel, Christian Sallaberry, Sébastien Laborie, Gilbert Tekli, Richard Chbeir: LinkedMDR: A Collective Knowledge Representation of a Heterogeneous Document Corpus. DEXA 2017: 362-377

3. Richard Chbeir, Yudith Cardinale, Pierre Bourreau, Khouloud Salameh, Nathalie Charbel, Lara Kallab, Chinnapong Angsuchotmetee, Gulben Calis: OntoH2G: A semantic model to represent building infrastructure and occupant interactions. KES-SEB 2018: 148-158.

4. Nathalie Charbel, Christian Sallaberry, Sébastien Laborie, Richard Chbeir: Hybrid Molecule-based Information Retrieval. ACM SAC 2019 - Accepted

1.6 Report Organization

The rest of this report is organized as follows.

Chapter 2 tackles the problem of **representing the collective knowledge embedded in a heterogeneous document corpus**. We first review existing standards and data models for document representation regardless of the application domain, then considering the AEC industry. We then describe our proposal within a semantic *tight coupling approach* which comprises (i) a formal definition of a *tightly coupled*

semantic graph, (ii) an introduction to *LinkedMDR* ontology and a description of its layers, and (iii) a *tight coupling algorithm* which provides the required pseudo code in order to generate the graph using the services of the Indexing Layer together with the knowledge provided by *LinkedMDR* in the Knowledge Representation Layer of *FEED2SEARCH*.

Chapter 3 tackles the problem of **IR over a heterogeneous document corpus** while providing the users with innovative search results. In other words, it addresses the challenge of searching for relevant information from the proposed tightly coupled semantic graph representing the corpus and providing query answers with meaningful context including both structural and domain-specific dimensions of a heterogeneous document corpus. We first review existing IR models including traditional and semantic models, and the variety of approaches and systems that implement them. We then propose a novel data structure for query answers which we call *Hybrid Molecules*, extracted from a tightly coupled semantic graph. We also detail on the novel *Hybrid Molecule-based Query Processing* Layer of *FEED2SAERCH*, with a focus on the Search and Ranking modules.

Chapter 4 presents the **experimental evaluation** study for the purpose of evaluating the contributions of Chapter 2 and Chapter 3 in the context of a domain-specific application, particularly the AEC industry. We first describe the implemented Java-based prototype with its two dedicated modules. The former, entitled *LMDR Annotator*, implements the Indexing Layer of *FEED2SEARCH*. It automatically annotates a heterogeneous document corpus and generates the *tightly coupled semantic graph*. The latter, entitled *HM Query Processor*, implements the Query Processing Layer of *FEED2SEARCH*. It exploits the generated semantic graph and provides a ranked list of Hybrid Molecule-based answers in response to a user's natural language query. We then describe a set of experiments conducted on construction projects, provided by Nobatek/INEF4. These experiments show promising results for an adoption in real-world applications.

Chapter 5 concludes this study and presents several **future research directions** that we are planning to explore afterwards.

Chapter 2

Towards a Collective Knowledge Representation of a Heterogeneous Document Corpus

"...when you connect data together, you get power."
- Tim Berners-Lee

The ever increasing need for extracting knowledge from heterogeneous data has become a major concern. One of the utmost challenges remains in representing this data, especially when it is dispersed among various types of interdependent documents generated from different sources. **This chapter tackles the problem of representing the collective knowledge embedded in a heterogeneous document corpus for it to be exploited by a Semantic Information Retrieval (SIR) system**. We propose a modular and generic semantic approach relying on a *tightly coupled semantic graph* where we associate semantics on two different dimensions: the content of the documents which depends on a domain-specific knowledge, and the structural and metadata information related to them. One of the major characteristics of this graph stands in the coupling information between hybrid components i.e., components of the two different dimensions. We generate such a graph based on a *tight coupling algorithm* and a backbone ontology which we call *LinkedMDR*. We introduce *LinkedMDR* as a novel multi-layered ontology. It offers core components modeling the documents, their relations and their metadata information at different granularity levels together with a pluggability feature that makes it adaptable to any domain-specific knowledge.

2.1 Introduction

Over the past two decades, modeling heterogeneous data has posed significant research [32, 56, 70, 92]. This has increasingly migrated into industrial applications aiming at providing the different actors intervening in a modern project with a powerful and intelligent search engine reducing their workloads, cognitive efforts and errors over heterogeneous document corpora. As mentioned in Chapter 1, this is particularly observed in the AEC industry. For instance, Fig. 2.1 illustrates some of the various heterogeneous documents[1] that are involved in a given construction project: d_1 and d_5 describe technical specifications of the building; d_2 describes thermal properties; d_3 describes acoustic properties; d_4 is an excerpt of a technical drawing related to the ground floor; and d_6 is an image depicting a material pattern used along the construction process.

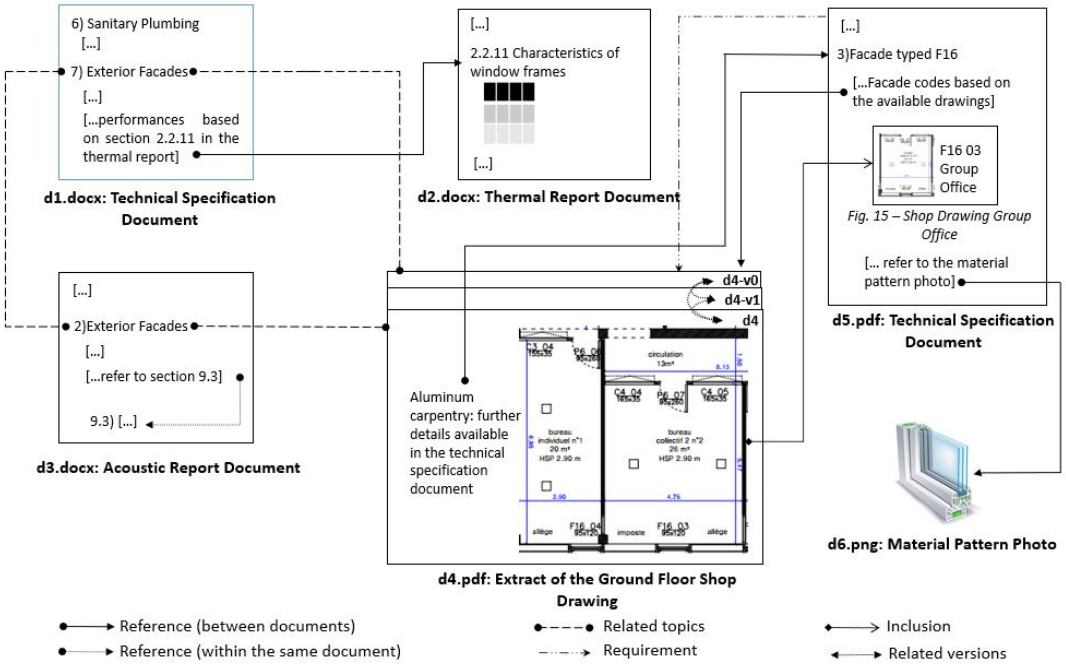


FIGURE 2.1 – Example of heterogeneous documents exchanged within a construction project.

In order to provide the basis for IR and search applications, several challenges arise for modeling such documents. In our context, we are particularly interested in the following challenges[2]:

- *Challenge 1.1: Representing various inter and intra-document links* - A document d_i has various relations among its components as well as with other documents. For instance, sections of d_1 and d_3 have related topics since they both describe the building's exterior facades. Section 2 of d_3 has an internal reference to

[1] For the sake of simplicity, we only present 6 documents. However, other documents could be also involved such as videos, audios and 3D drawings.

[2] Although other challenges exist (e.g., data confidentiality and security), they remain out of scope.

its section *9.3*. d_5 has an external reference to the image d_6 and the technical drawing d_4, which itself has several versions.

- *Challenge 1.2: Representing metadata on the documents, their content, and their structure* - There are metadata descriptors providing useful generic information on the document as a whole entity (e.g., a given version of the technical drawing d_4), on its content (e.g., a description on the content of the image d_6), and on its structure on different depth levels (e.g., section *7* of d_1, figure *15* of d_5).

- *Challenge 1.3: Handling advanced information semantics* - Semantics should be associated with the content of the documents, their relations, and their metadata information on the different granularity levels. This helps to smartly reason over the collection of documents. For instance, one needs to locate, on the finest granularity levels, relevant information regarding the exterior facades which is contained in several documents: section *7* of d_1, section *2.2.11* of d_2 referenced by d_1, the related technical drawing d_4, etc.

- *Challenge 1.4: Handling documents multimodality* - Coping with multimodality means handling multimedia content but also the different formats the documents could have. For instance, d_1, d_2 and d_3 are Word documents, d_4 is a CAD in a PDF format, d_5 is a PDF document and d_6 is a PNG image.

- *Challenge 1.5: Ensuring extensibility* - The information may evolve over time. This may involve new document link types, new vocabulary, new document formats, new media types, etc. For instance, an extension of the scenario illustrated in Fig. 2.1 would integrate an audio file that analyzes the noise impact before and after the cladding of the facades.

In the literature, several works have been undertaken to define standards [29, 31, 97, 98] and data models [4, 13, 14, 37, 70, 83] for document representation. On the other side, other domain-oriented works, such as building-oriented standards [12, 16, 41] and data models [18, 57, 58, 59] have taken up the challenge of modeling building data and managing related documents [68, 89]. To the best of our knowledge, none of the existing works have tackled the above challenges combined.

In this context, our aim is to provide a clear and a complete picture of the collection of heterogeneous documents w.r.t. the identified issues, so it could be exploited by an IR system. Thus, in this chapter, we present a modular and generic *tight coupling approach* where, given a heterogeneous document corpus and an external domain-specific knowledge, it generates *a tightly coupled semantic graph* representing the collective knowledge embedded in that corpus based on two different dimensions: the domain-specific information that reflects the content of the documents, and the structural and metadata information describing them. We thus introduce *LinkedMDR* [20], a multi-layered ontology, which adapts to any domain application

and provides the required infrastructure to generate this graph, including the capability of: (i) representing various inter and intra-document relations, (ii) representing metadata information at different levels of precision while relying on the existing standards, (iii) handling advanced information semantics on the two different dimensions, (iv) handling multimodal documents, and (v) ensuring extensibility. We also present a *tight coupling algorithm* which details the procedure of generating such a graph.

The contributions of this chapter are summarized by:

- The *Tight Coupling Approach*, which provides a general overview of our proposal;
- The *Tightly Coupled semantic graph*, which is formalized on the basis of the two identified dimensions of a heterogeneous document corpus;
- *LinkedMDR* [20], which is the backbone ontology to generate this graph;
- The *Tightly Coupled Algorithm*, which provides a means to implement our proposal.

The remainder of this chapter is organized as follows. Sect. 2.2 reviews existing standards and data models for document's structure and content description as well as for the building data in the AEC industry. Sect. 2.3 describes the proposed *tight coupling approach* which comprises the *tightly coupled semantic graph*, *LinkedMDR* ontology and the underlying *tight coupling algorithm*. Sect. 2.4 concludes the chapter.

2.2 Related Work

In this section, we review existing standards and models addressing (i) metadata for document representation in general, and (ii) building data in particular. This literature review is based on academic researches and industrial solutions as well.

2.2.1 Metadata Standards and Data Models for Document Representation

The metadata standards and data models for document representation are domain independent. They can be divided into four different categories: (1) single media-based standards which handle one type of media content, (2) multimedia standards which handle multiple types of media content, (3) ontology and knowledge-based models which build on existing multimedia standards to provide more semantics, and (4) other models which build on traditional ways of representing the data neglecting semantics of the information.

2.2.1.1 Single Media-based Standards

We consider single media-based standards those basically describing only text or image contents. Renowned standards dedicated for audio and video are, in general,

multimedia-based as they involve at least two media content types in their descriptions[3]. Although there exist many other standards in this category, we present an example on the most commonly used standard for image description and another one for text description:

EXIF - The Exchangeable Image File (EXIF) format is a widely used standard for describing digital images [31]. It mainly supports a set of tags related to image data structure (e.g., width, height, pixel composition), version, data characteristics (e.g., color space information), configuration (e.g., image compression mode), user information (e.g., user comments), date and time, recording offset (e.g., bytes of JPEG data), picture-taking conditions (e.g., exposure time, brightness), GPS (e.g., latitude, altitude) and many other tags (e.g., image title, software used, creator, copyright).

TEI - The Text Encoding Initiative (TEI) [98] is a commonly adopted text-driven descriptive standard. It is based on the eXtensible Markup Language (XML) format and provides a way to describe information and meta-information within a textual document. TEI offers a representational form of a text as well as a set of semantically rich elements that could improve IR. There is no formal grouping of TEI elements. However, it is possible to classify them along the following categories: the structural elements (e.g., chapters, sections, paragraphs, lists, items, page break, table, cross reference links), the highlighting elements (e.g., highlighted text, italic, bold), the logical and semantic elements (e.g., title, name, measure, date, address, abbreviation, emphasis), the analytical elements (i.e., additional information, e.g., notes, corrections, indexes), and the figures and graphics[4].

2.2.1.2 Multimedia-based Standards

Although there exist many other standards in this category, we describe the most adopted ones in the literature:

DC - The Dublin Core (DC) Metadata Initiative[5] is a metadata standard for describing a wide range of online multimedia resources. The DC Metadata Element Set consists of 15 Elements describing the content of a document (e.g., title, description, etc.), the intellectual property (e.g., creator, rights, etc.), and its instantiation (e.g., date, format, etc.) [103]. This standard also offers a set of qualifiers which aim to modify the properties of the DC statements [29].

MPEG-7 - The Multimedia Content Description Interface (MPEG-7) is an ISO/IEC standard developed by MPEG (Moving Picture Experts Group) [97]. Its aim is to provide a rich set of complex standardized tools describing low and high level features

[3] For instance, ID3 standard describes digital audio files which can contain, in addition to the audio track, related text and/or graphical information.

[4] `http://teibyexample.org/modules/TBED01v00.htm`

[5] `http://dublincore.org/specifications/`

while also covering different granularities of data (e.g., collection, image, video, audio, segment) and different areas (content description, management, organization and navigation). The three main standardized components of MPEG-7 are: Descriptors (Ds), Description Schemes (DSs) and Description Definition Languages (DDL). While the MPEG-7 Ds are representations of features, the MPEG-7 DSs support complex descriptions and specify the structure and the semantics of the relationships among its constituents: Ds and DSs [84]. The MPEG-7 DDL is a standardized language based on XML schema. It allows the extension of existing Ds and DSs as well as the introduction of new components for specific domains [50].

2.2.1.3 Ontology and Knowledge-based Models

There is often the need to combine several standards in order to meet the requirements of complex multimedia applications [87]. Many initiatives have been taken for the purpose of building multimedia ontologies and knowledge bases, such as [4, 70, 83, 102], or transforming existing formats into ontologies, such as [37]. The aim of these studies is to bridge the gap between low level features with automatically extractable information by machines and high level human interpretable features of the same information [92]. The following are examples on well-known ontologies and knowledge bases:

COMM - The Core Ontology for MultiMedia (COMM) [4] is an MPEG-7 compliant ontology, designed to facilitate multimedia annotations. Although it is not aligned with the XML Schema of the MPEG-7 standard, COMM covers its most important parts, specifically the structure of the media and the content of multimedia documents, while providing more formal semantics. It is designed based on the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) as a foundational ontology, and two Ontology Design Patterns (ODPs): the Descriptions & Situations (D&S) and the Ontology of Information Objects (OIO) formalizing contextual knowledge and information objects respectively. A Java API is provided in order to translate the objects of the MPEG-7 classes into instances of the COMM concepts.

M3O - The Multimedia Metadata Ontology (M3O) [83] aims to provide abstract pattern-based models for multimedia metadata representation. It is centered on the formal upper-level ontology DOLCE+DnS Ultralight (DUL) of which it provides three specialized patterns (Description and Situation (D&S), Information and Realization, and Data Value) together with two other patterns (Annotation and Decomposition). M3O is aligned with several ontologies such as COMM, Media Resource Ontology and the EXIF standard.

MediaOnt - The Media Resource Ontology (MediaOnt) is developed by the W3C Media Annotation Working Group [102] for describing media resources, specifically images, video and audio fragments. It is made of a core vocabulary comprising a

set of descriptive properties covering identification, creation, content description, rights, relational, distribution, fragments and technical properties. MediaOnt is subject to multiple alignments with several existing multimedia metadata, such as MPEG-7, DC and EXIF. The main purpose of the mappings is to ensure interoperability among the most commonly adopted metadata formats on the web describing media resources [92].

Mepg-7 Rhizomik - The Mpeg-7 Rhizomik ontology is the first complete MPEG-7 ontology designed based on one-to-one automatic mapping of the entire standard to Web Ontology Language (OWL) [37] using XSD2OWL[6]. The XML schema to OWL mapping is complemented with a mapping of XML metadata instances to Resource Description Framework (RDF) in order to generate semantic-enabled representations of the input metadata instances and thus facilitate data integration when several sources are available. The Mpeg-7 Rhizomik ontology have been used as an upper-level ontology for other domain ontologies, such as music ontology.

OntoText Data Model - The OntoText company[7] provides proprietary solutions for data analytics for global businesses. These solutions are based on semantic data models in order to (i) represent large amount of information coming from diverse data sources, (ii) create meaningful connections between structured and unstructured data, and (iii) obtain valuable insights by reasoning over the semantic graphs. The basic approach behind the construction of such graphs is not to rely on a common data model but to automatically extract Named Entities and rules that create relations between them [70]. These entities basically cover Person, Organization, Location from renowned knowledge graphs such as DBpedia[8], Wikidata[9] and Schema.org[10].

2.2.1.4 Other Models

Although there exist many other non ontology or knowledge-based models, we present the following two models as examples on a renowned data model for a specific document format (PDF) and a renowned generic data model for multimedia content and various document formats:

XCDF Data Model - Problems such as over-segmentation, additional noisy information, and lack of structure preservation produced by PDF generators are often associated with the PDF format. Great effort has been put into overcoming these issues [13]: XCDF is a canonical format which purpose is to represent the results of the

[6] http://rhizomik.net/html/redefer/#XSD2OWL
[7] https://ontotext.com/
[8] https://wiki.dbpedia.org/
[9] https://www.wikidata.org/wiki/Wikidata:Main_Page
[10] https://schema.org/

physical structure extraction of the PDF documents in a single and structured format. It is based on the XML format and its DTD provides basic elements for general document representation (page, fonts, image, graphic, textblock, textline and token).

LINDO Meta Model - In the context of distributed multimedia information systems, such as the video surveillance applications that use different indexing engines, interoperability problems arise when metadata of different formats are combined together. The LINDO project (Large scale distributed INDexation of multimedia Objects)[11] takes up the challenge of handling different metadata standards within its distributed information system, such as DC, EXIF and MPEG-7. Thus, the authors in [14] define a unified XML-based metadata model that encapsulates these standards based on two levels: general metadata information describing the entire document and metadata related to multimedia contents (image, text, video and audio).

2.2.2 Building-Oriented Standards and Data Models

The building-oriented models are limited to applications in the AEC industry. They can be divided into three different categories: (1) standards which model the building data, (2) ontology-based models which build on existing standards to provide more semantics, and (3) other models which build on traditional ways of representing the building data neglecting semantics of the information.

2.2.2.1 Standards

Although there exist many other standards in this category, we present the widely known standards:

IFC - The IFC (Industry Foundation Classes) [16] is one of the most recent and widely used information exchange standards in the building and construction industry. It is developed by buildingSMART alliance[12] in the context of the openBIM approach for collaborative design, realization and operation of buildings. The main purpose is to define an open and common data schema that is independent of the different software applications. This standard is object-oriented [9]. It involves building and construction objects (including physical components, spaces, systems, processes and actors) and relationships between them [51]. IFC specifications are written using the EXPRESS data definition language[13], but they are also published in XSD (XML Schema Definition).

[11] `https://itea3.org/project/lindo.html`
[12] `https://www.buildingsmart.org/`
[13] Defined as ISO10303-11 by the ISO TC184/SC4 committee

COBie - The Construction Operations Building information exchange (COBie) [12] is a non proprietary international standard for the exchange of non geometric building information. It was initially led by the Engineer Research and Development Center, Construction Engineering Research Laboratory of the U.S. Army, Corps of Engineers. It is also implemented by the buildingSMART international as a subset of the IFC [16], specifically as a Model View Definition (MVD) [17] of this standard with simplified information needed by facility managers. It focuses on the operations, maintenance and asset management information related to equipment and spaces. The COBie data is serialized in EXCEL spreadsheets or IFC file formats such as XML and STEP files. For the purpose of providing a specification of the way that information can be transferred between the spreadsheets and IFC versions, the COBie standard is integrated as part of the US National BIM Standard (NBIMS)[14] version 2.

gbXML - The Green Building XML standard (gbXML) [41] is an open XML schema, originally submitted by Green Building Studio Company[15], in order to ease the transfer of BIM data between disparate building design software tools for energy performance analysis. The information mostly covered by the gbXML standard is related to spaces and surfaces [107] along with other elements leveraging the overall energy consumption of the building (e.g., schedule of use, thermostat temperature set-points, internal loads, materials and constructions). The core idea behind it is to describe multiple buildings that are somehow related and located in the same climate region.

2.2.2.2 Ontology-based Models

Although there exist many other ontologies in this category, we present the most referenced by the literature:

ifcOWL - The ifcOWL [18] is the representation of the IFC standard [16] into ontology by converting the EXPRESS schema of the standard into OWL. The major goal is to migrate into SW technologies to support data interoperability, flexibility and extensibility within information system applications in the construction industry. In this direction, there have been several approaches for the EXPRESS-to-OWL conversion, such as [2, 6, 88, 96]. Among all approaches, Pauwels and Terkaj [74] propose a recommendable and usable ifcOWL ontology. It is semantically closer to the original IFC Schema Standard in comparison to alternative approaches. The ontology encapsulates 1230 Classes, 21306 Axioms, 1578 Object Properties and 5 Data Properties.

[14] `https://www.nationalbimstandard.org/`
[15] `http://www.gbxml.org/`

BOT - The Building Topology Ontology (BOT) [58] is a modular ontology covering core concepts of the building and three methods for extending it with domain specific ontologies (such as ontologies modeling information related to geographical location, sensor data, domotics, and construction data). The main purpose is to reduce redundancies caused by overlapping data found in these ontologies, which is currently violating the W3C best practice rules. Further, this approach helps the development of distributed ontologies, which have several advantages over a one-size fits all, such as interoperability and simple reuse. The BOT ontology comprises 7 key Classes (Building, Element, Interface, Site, Space, Storey, and Zone), 14 Object properties linking them and 1 Data Property linking a Zone or an Element to an IRI that identifies its 3D Model. As for the linking methods, it comes down to (i) extending the core classes with subclasses of more specialized ontologies, (ii) defining equivalent classes, and (iii) establishing typed links between instances [79].

PRODUCT - The Building Product Ontology (PRODUCT) [57] is another modular ontology developed along the same lines as BOT [58], however for describing a product i.e., an article or substance that is manufactured or refined for sale. It encapsulates one main Class (Product) and one main Object Property (Aggregates). The latter represents a simple decomposition tree of products. A custom product of more specific ontologies can be specified using this class and object property (such as the Element Class of the BOT ontology being a subClass of Product).

OPM - The Ontology for Property Management (OPM) [59] is an ontology for describing temporal properties that are subject to changes as the building design evolves. It is made of 8 Classes, 5 Object Properties, and 2 Data Properties modeling the building properties and their states along the project life-cycle.

2.2.2.3 Other Models

Although there exist many other non ontology-based models, we present two well-known of them here: the first one in the international market and the second one in France:

Newforma Data Model - The Newforma company provides proprietary solutions for PIM [68] mainly dedicated for architects, engineers, contractors and owners. The main purpose of these solutions is to index all information in a given project including unstructured data (such as emails, construction related files of different formats). This helps the different users in organizing and managing project data, and collaborating with the different team members. The underlying data model represents general metadata information, syntactic words extracted from the content of the documents, and keywords manually added by the users.

Kroqi Data Model - Kroqi [89] is a free collaborative platform developed by the Scientific and Technical Center Building[16] (CSTB) as a tool for the digital transition in the AEC industry to reduce the gap of the application of the BIM in large enterprises and small to medium-sized enterprises. Kroqi comprises several modules including the document management module. The latter provides several features, yet we focus on the document representation feature. The underlying data model represents general metadata describing the documents and tags manually added by the users.

2.2.3 Discussion

We evaluated the existing standards and models based on the challenges previously mentioned in Sect. 2.1. The results are depicted in Table 2.1 where we used the following symbols to evaluate a given standard or data model: "✓" to express an exhaustive coverage, "✗" to express a lack of coverage, and "**Partial**" to express a partial coverage.

- *Inter/Intra-document Relations (Challenge 1.1)*: In general, there is no existing standard or data model capable of fully representing inter and intra-document relations. For instance, TEI [98] provides $\langle ref \rangle$ element which is only limited to cross references. DC [29] comprises a set of relation qualifiers (e.g., *References*, *isVersionOf*) between different resources but excludes intra-document relations (e.g., reference between parts of the same resource) and more complex inter-document relations (e.g., a spatial relation where a specific part of a resource is contained in another resource). Mpeg-7 [97] and multimedia ontologies (e.g., [4, 37, 83, 102]) provide capabilities of linking only audiovisual descriptors (e.g., a semantic relation between two objects of a video segment). The LINDO data model [14] covers only spatio-temporal relations between objects describing the content of multimedia elements (e.g., *Localization* relation). The Ontotext knowledge-based graph [70] is capable of describing relations between entities representing the content of the same or different web pages (e.g., a semantic relation between an entity representing an organization and another entity representing its CEO). Yet, this is still limited as it does not cover other relations (e.g., references between structural metadata). This is also the case with building-oriented standards (e.g., [16, 12, 41]) and ontologies (e.g., [18, 57, 58, 59]) which only describe relations between entities representing the building data. The other building-oriented data models (e.g., [68, 89]) provide inter-document relations limited to document versions since they are conceptually designed towards document management solutions.

- *General, Content-based, and Structural Metadata (Challenge 1.2)*: Looking at the representation of metadata (generic, content-based, and structural metadata),

[16] `http://www.cstb.fr/`

TABLE 2.1 – Evaluation of the existing standards and data models w.r.t. the identified challenges.

	Challenge 1.1	*Challenge 1.2*			*Challenge 1.3*	*Challenge 1.4*	*Challenge 1.5*
	Inter/Intra Document Relations	**Generic Metadata**	**Content Description**	**Structural Metadata**	**Advanced Semantics**	**Multimodality**	**Extensibility**
EXIF [31]	✗	✓	✗	✗	✗	✗	Partial
TEI [98]	Partial	✓	Partial	Partial	✗	✗	Partial
DC [29]	Partial	✓	✗	✗	✗	✓	Partial
MPEG-7 [97]	Partial	✓	✓	Partial	✗	Partial	Partial
COMM [4]	Partial	✓	✓	Partial	Partial	Partial	Partial
M3O [83]	Partial	✓	✓	Partial	Partial	Partial	Partial
MediaOnt [102]	Partial	✓	✓	✗	Partial	Partial	Partial
MPEG-7 Rhizomik [37]	Partial	✓	✓	Partial	Partial	Partial	Partial
Ontotext Data Model [70]	Partial	✓	✓	✗	Partial	Partial	Partial
XCDF Data Model [13]	✗	✗	✗	Partial	✗	✗	Partial
LINDO Meta Model [14]	Partial	✓	✓	Partial	✗	✓	Partial
IFC [16]	Partial	✓	Partial	✗	✗	✗	Partial
COBie [12]	Partial	✓	Partial	✗	✗	✗	Partial
gbXML [41]	Partial	✓	Partial	✗	✗	✗	Partial
ifcOWL [18]	Partial	✓	Partial	✗	Partial	✗	Partial
BOT [58]	Partial	✗	Partial	✗	Partial	✗	Partial
PRODUCT [57]	Partial	✗	Partial	✗	Partial	✗	Partial
OPM [59]	Partial	✗	Partial	✗	Partial	✗	Partial
Newforma Data Model [68]	Partial	✓	Partial	✗	✗	✓	Partial
Kroqi Data Model [89]	Partial	✓	Partial	✗	✗	✓	Partial

there is no existing standard or data model providing a full coverage of the three aspects combined. Although, the majority fully cover generic metadata, they do partially cover content description and often lack structural metadata description. For instance, EXIF [31] and DC [29] are limited to metadata information describing a document as a whole entity. The XCDF data model [13] only focuses on structural metadata describing the decomposition of PDF documents. TEI [98] provides a rich set of content description and structural metadata focusing however on the text. On the other hand, the multimedia-based standards (e.g., [97]), ontology and knowledge-based models (e.g., [4, 37, 70, 83, 102]), and other data models (e.g., [14]) provide a wider coverage for the description of multimedia content while failing in the description of structural metadata, especially those related to the text (e.g., the page encapsulating a given textual content). The building-oriented standards (e.g., [16, 12, 41]) and ontologies (e.g., [18, 57, 58, 59]) are not conceived to describe metadata information. The data models behind building-oriented document management solutions (e.g., [68, 89]) are however capable of representing general metadata on the documents. Although they do not provide metadata on the content of the documents, they allow the users to add tags that might describe the content of the documents.

- *Advanced Semantics (Challenge 1.3)*: As for the advanced semantics capabilities, none of the existing standards or data models contains the required knowledge for their associated systems to reason over the documents together with their content and their structure. However, the ontology and knowledge based models, whether representing multimedia document's content (e.g., [4, 37, 70, 83, 102]) or the building data (e.g., [18, 57, 58, 59]), partially integrate some advanced semantic aspects which are limited to reasoning capabilities over the content of the documents.

- *Multimodality (Challenge 1.4)*: The multimodality aspect is covered by few of the existing standards and data models. For instance, DC [29] is capable of describing any resource regardless of its type, its media content types, and its serialization technology. The LINDO data model represents multimedia contents that could be encoded in different formats. The building-oriented data models behind the industrial solutions for document management (e.g., [68, 89]) are basically designed to support multimodality (e.g., emails, textual documents of different formats, images of different formats, 3D files). The multimedia standards, ontologies and knowledge-based models [4, 37, 70, 83, 97, 102] naturally handle multimedia contents, yet they partially cover the multimodality aspect since they are not further capable of representing textual documents in their different forms of serialization.

- *Extensibility (challenge 1.5)*: We consider all the reviewed standards and data

models as partially extensible i.e., the evolution of the model itself is possible, yet following narrow directions. For instance, the nature of XML-based standards and data models makes them partially extensible as they could support new elements, however describing their particular coverage. For instance, defining new DSs in Mpeg-7 [97] is possible, which makes it somehow extensible. Yet, defining more complex elements such as inter and intra-document links is still out of its scope.

To sum up, to our knowledge, there is currently no available standard or data model that addresses all the challenges and answers all the requirements we described in Sect. 2.1. A naive solution could be the simple combination of the most convenient standards or data models for document representation, and the building data. However, this yields interoperability issues since they rely on different semantics, structures and formats. We believe that ontologies provide a reliable and efficient means to resolve these problems and support SIR from heterogeneous data [44]. Thus, a more complex solution could be also the alignment of the best suited ontologies. Nonetheless, this would partially solve their current limitations as shown in Table 2.1. Furthermore, we aim to provide a modular and generic solution that is able to adapt to particularities of a given domain (e.g., the AEC industry) but also interpolated in any other domain.

2.3 Tight Coupling Proposal

Our aim is to provide powerful knowledge representation capabilities over the various data embedded in a heterogeneous document corpus for it to be exploited by any domain-oriented IR application. In this context, we need to leverage semantics on two different dimensions of a document corpus: the content of the documents which depends on the knowledge of a given application domain, and the structural and metadata aspect describing the documents regardless of that domain. Once semantics are handled, we propose to couple domain-specific information with its related structural and metadata information.

2.3.1 Overview

We provide a modular and generic semantic approach capable of generating a semantic network called *tightly coupled semantic graph* describing the collective knowledge of any heterogeneous document corpus in view of the limitations of current standards and data models (See Sect. 2.2). Our proposal is depicted in Fig. 2.2, which corresponds to the Indexing and Knowledge Representation layers of *FEED2SEARCH* (See Sect. 1.4.1).

In our approach, we rely on *ontologies* since they are proven powerful in: (1) conceptualizing heterogeneous data, (2) handling interoperability between vocabularies

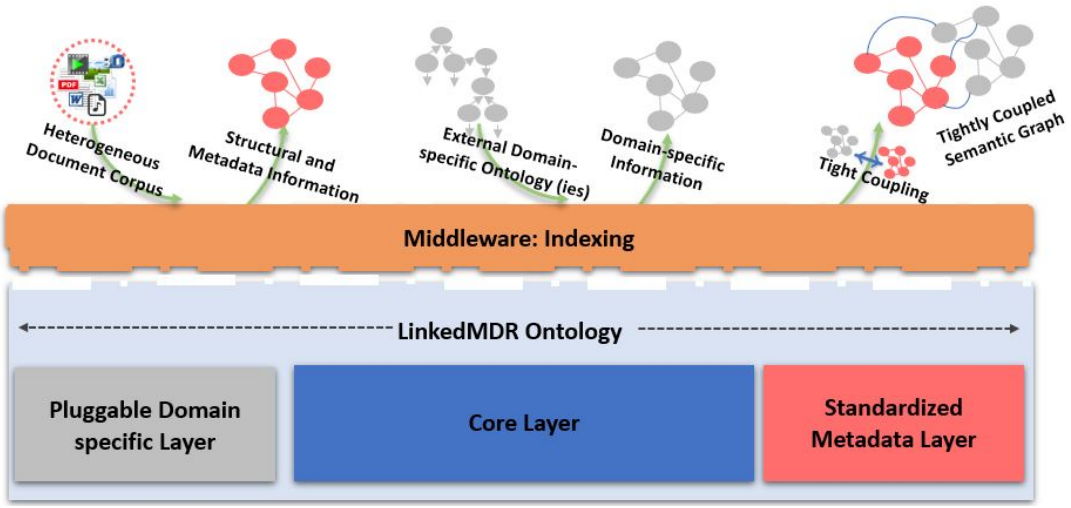


FIGURE 2.2 – Overview of the our Tight Coupling Approach

of different standards or data models, (3) covering lexico-syntactic and semantic aspects, (4) providing advanced reasoning capabilities, (5) ensuring extensibility, and (6) supporting SIR [91].

We propose *LinkedMDR*, a novel multi-layered ontology, which provides the infrastructure for the generation of a tightly coupled semantic graph. A middleware is required in order to benefit from the different existing services, technologies and APIs, and translate the generated information into *LinkedMDR* instances, thus progressively building the semantic network. Given a heterogeneous document corpus, one or multiple external domain-specific ontologies describing the knowledge embedded in the corpus, *LinkedMDR* ontology and its dedicated middleware:

- Structural and metadata instances describing the heterogeneous document corpus are generated using concepts, relations, and semantic rules of the core layer and the standardized metadata layer of *LinkedMDR*,
- Domain-specific instances describing the content of the heterogeneous document corpus are generated using concepts, relations, and semantic rules provided by the core layer of *LinkedMDR* and its pluggable domain-specific layer which adapts to the external domain-specific ontology(ies),
- Relations between instances describing the structural and metadata information and those describing the domain-specific information are generated using the semantic knowledge provided by the core layer of *LinkedMDR* and a tight coupling algorithm.

In the following, Sect. 2.3.2 details on *LinkedMDR* ontology and its three layers, Sect. 2.3.3 describes a generated tightly coupled semantic graph together with the underlying tight coupling algorithm.

2.3.2 LinkedMDR Ontology

We introduce a novel multi-layered ontology, entitled *LinkedMDR*[17] [20], which stands for ***Linked*** ***M***ultimedia ***D***ocument ***R***epresentation. The main purpose of this ontology is to provide the infrastructure to model the knowledge embedded in a heterogeneous document corpus of any domain-specific application. *LinkedMDR* is made of three main layers:

- The *Core* layer serving as a backbone and a mediator among the different layers,
- The *Standardized Metadata* layer linking descriptors of existing standards in metadata representation,
- The *Pluggable Domain-Specific* layer that can adapt to any domain-specific ontology.

The multi-layering of *LinkedMDR* ensures its genericity and extensibility. Fig. 2.3 shows an example of *LinkedMDR* application to the AEC industry using DC [29], TEI [98] and MPEG-7 [97] standards for the standardized metadata layer and the ifcOWL [18] ontology for the pluggable domain-specific layer.

2.3.2.1 Core Layer

This layer introduces new concepts and relations that are either not covered or partially covered by existing standards (See Sect. 2.2), mainly:

- Concepts that model the global composition of a given document and the metadata properties associated to it (i.e., *Document*, *Media*, *MediaComponent* and *DocumentProperty*). This allows the description of various characteristics on the different granularity levels of the document.
- *Object* concept which abstracts *Document*, *Media* and *MediaComponent* so as to define on the *Object* common characteristics that are then inherited by these concepts.
- A rich set of relations associated to the *Object* concept (i.e., *Part-Whole*, *Semantic*, *Temporal* and *Spatial* relations) to define possible relations between any two document components (i.e., *Document*, *Media* and *MediaComponent*) or associated to a specific document component (e.g., *Order* and *Syntactic* relations associated to the *Media* concept) to define particular relations between two entities of the same type.
- Concepts that are equivalent to other concepts of adjacent layers (e.g., *Text* and *Image* describe respectively text and image concepts of the standardized metadata layer). This ensures easy and flexible alignments with similar concepts of adjacent layers that are defined by external standards or data models.

[17] Available at `http://spider.sigappfr.org/linkedmdr/`

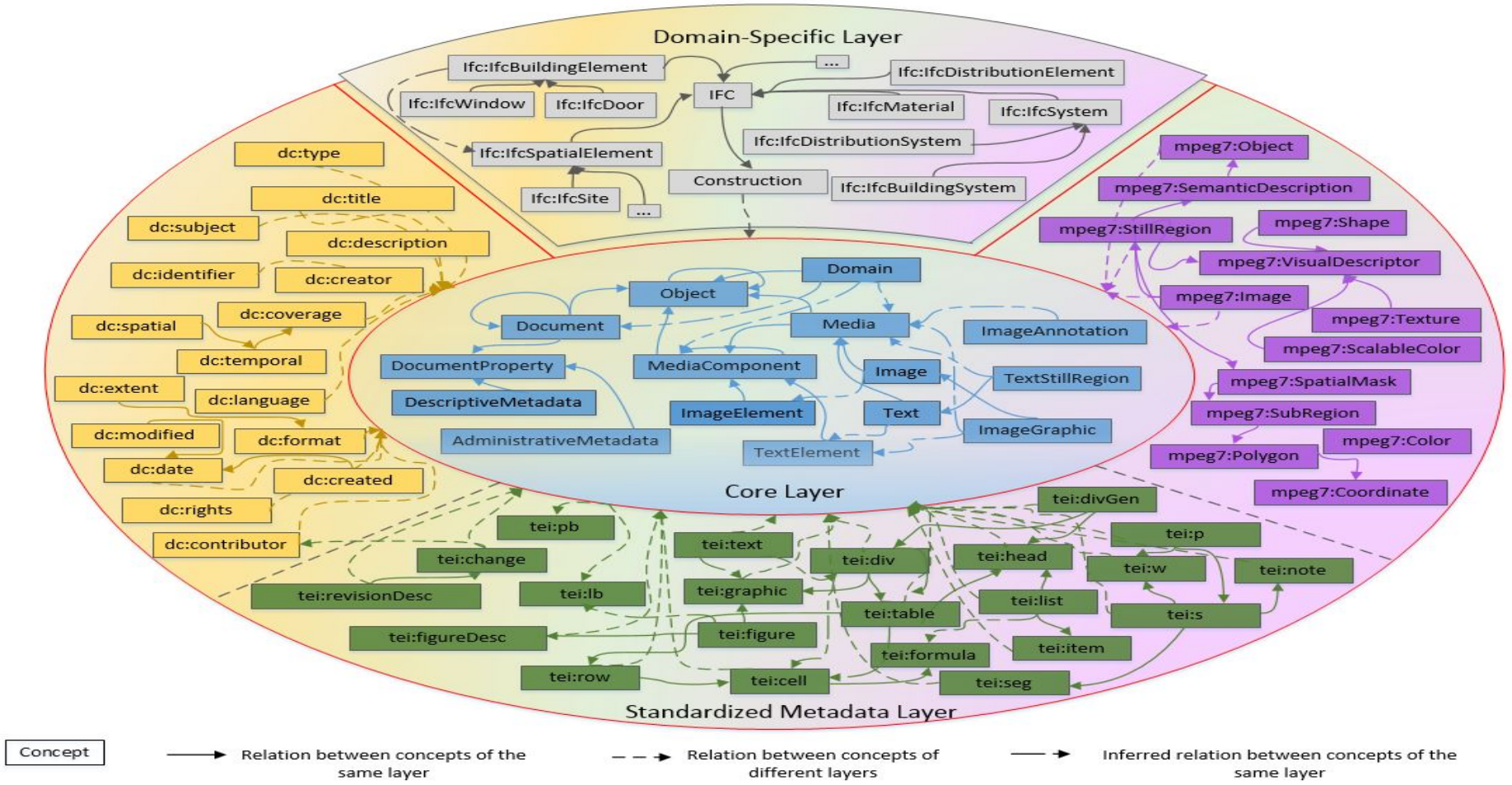


FIGURE 2.3 – Overall schema architecture of *LinkedMDR* ontology.

- Concepts that subsume other concepts of adjacent layers (e.g., *DescriptiveMetadata*, *AdministrativeMetadata*, *Text*, *Image*, *TextElement* and *ImageElement* are super concepts generalizing descriptors of the standardized metadata layer; *Domain* is a super concept generalizing an entity describing a particular domain of the pluggable domain-specific layer). This allows to define core concepts' characteristics that can be then inherited by other concepts of adjacent layers regardless of the underlying external standard or data model.
- Concepts that extend some descriptors of existing standards in the standardized metadata layer (e.g., *TextStillRegion* describes text metadata inside an image, *ImageGraphic* describes image metadata inside a text, and *ImageAnnotation* annotates an image with another structured image or structured text). This creates links between descriptors of different existing standards through core concepts of *LinkedMDR* and ensures the reuse of these descriptors while adapting them to the requirements of our context.

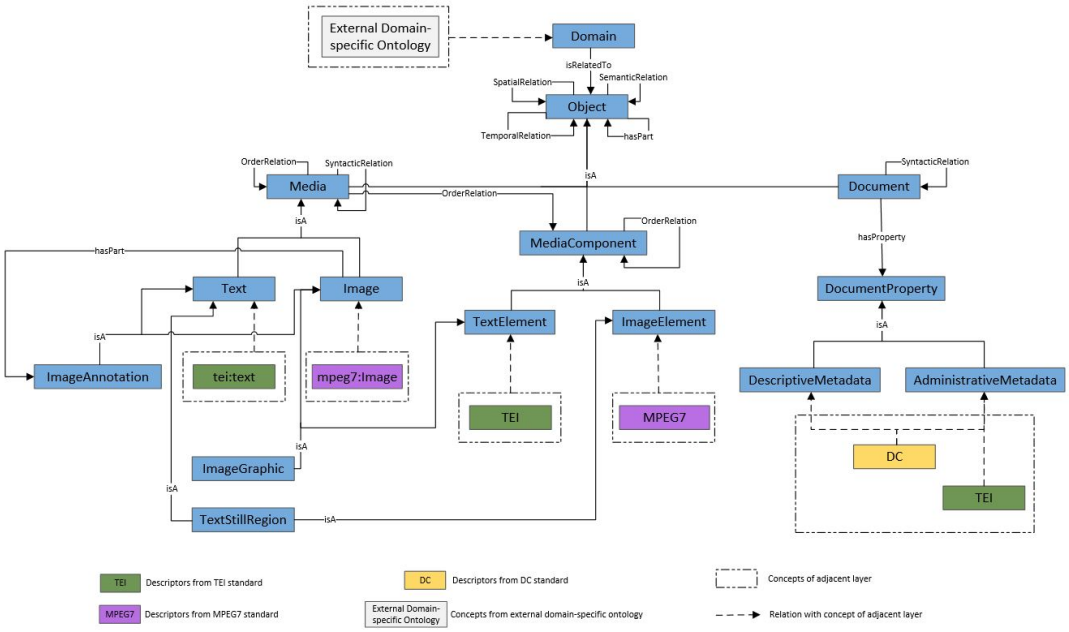


FIGURE 2.4 – Overview of *LinkedMDR* Core layer.

Fig. 2.4 shows the overall schema of the core layer and its connections to elements of other layers. A detailed documentation on *LinkedMDR* core concepts and relations is available online at: `http://spider.sigappfr.org/linkedmdr/documentation/`.

2.3.2.2 Standardized Metadata Layer

This layer builds upon metadata information defined by existing standards. Its main purpose is to reuse the most convenient descriptors representing general metadata of the document, textual and multimedia components while adapting them to the previously mentioned requirements (See Sect. 2.1). As an example, we select DC [29], TEI [98] and MPEG-7 [97] as they present a rich set of descriptors satisfying the three

above categories of descriptors respectively (See Sect. 2.2). Although our selection is rational, we still leave flexibility for the use of other standards in the future for later upcoming versions of *LinkedMDR*.

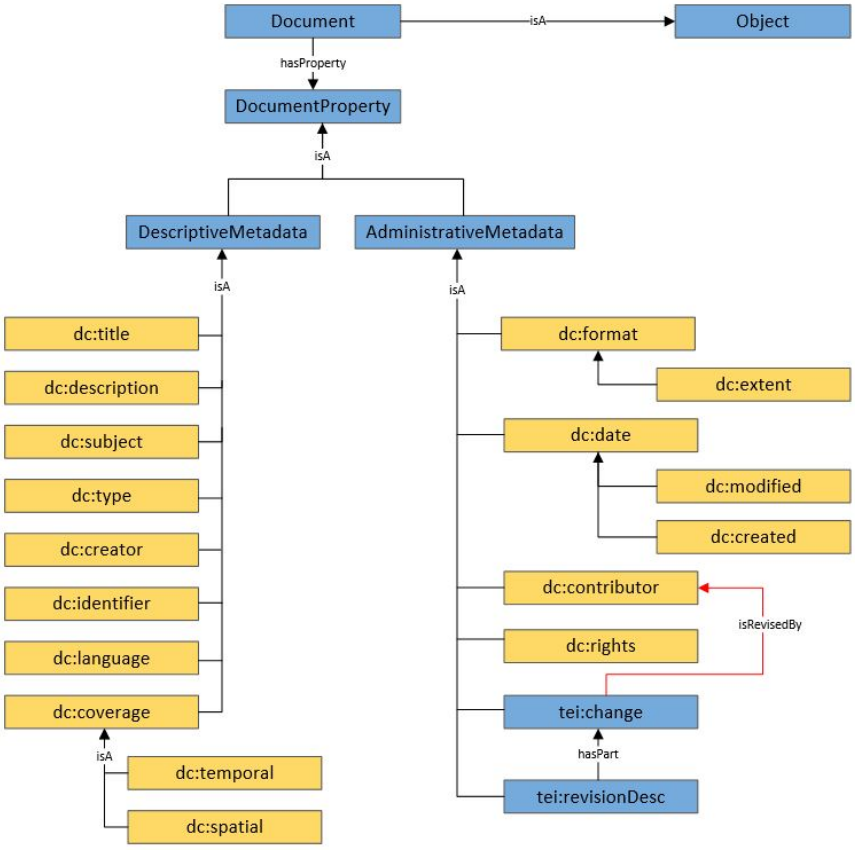


FIGURE 2.5 – Example of *LinkedMDR* Standardized Metadata sub-layer dedicated to DC standard.

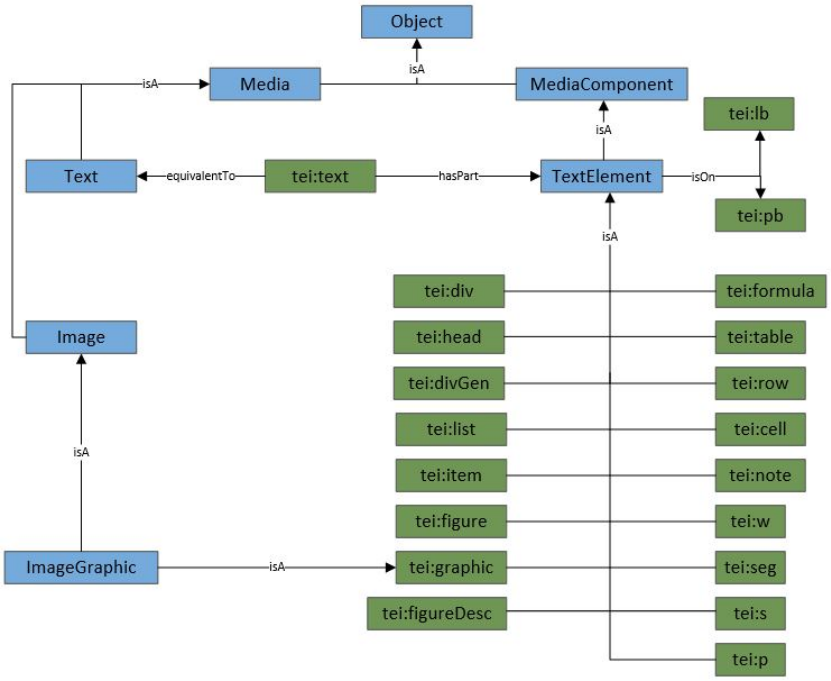


FIGURE 2.6 – Example of *LinkedMDR* Standardized Metadata sub-layer dedicated to TEI standard.

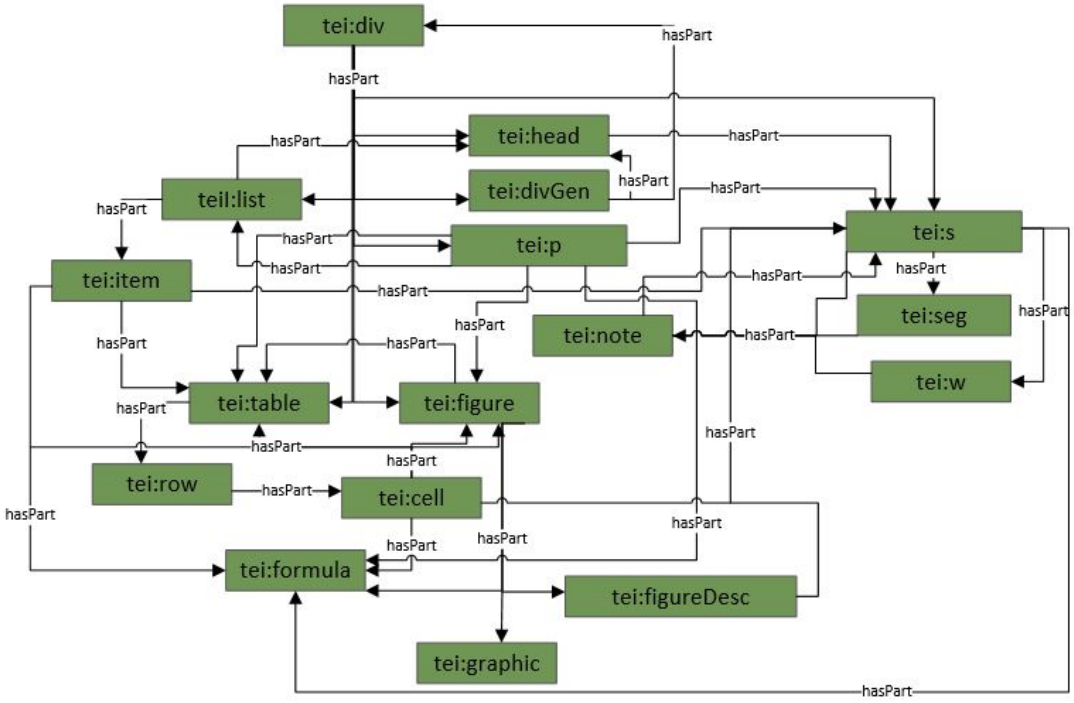


FIGURE 2.7 – Extract of relations between concepts from TEI standard in the corresponding *LinkedMDR* Standardized Metadata sub-layer.

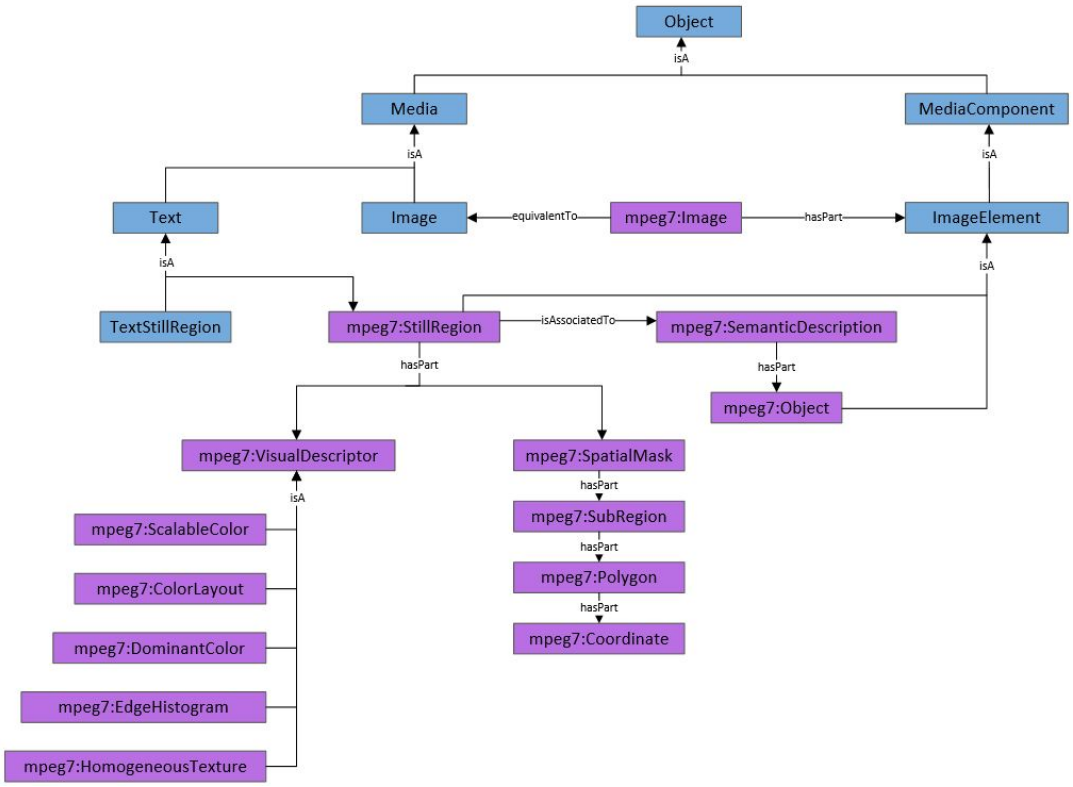


FIGURE 2.8 – Example of *LinkedMDR* Standardized Metadata sub-layer dedicated to MPEG-7 standard.

So far, this layer is divided into three sub-layers, each dedicated to a standard. The first corresponds to DC and comprises metadata information of the document in general (See Fig. 2.5). The second is related to TEI structural text metadata (See Fig. 2.6 and Fig. 2.7). The third contains metadata information that describes multimedia data following the MPEG-7 standard. For the sake of simplicity, we only present the image metadata with different levels of precision, its related visual features and semantic descriptors (See Fig. 2.8). This layer also involves relations between its sub-layers. For instance, *isRevisedBy* relation (the red highlighted link in Fig. 2.5) is added in order to link the concept *tei:Change*, which describes a set of changes made during the revision of a document, to the corresponding concept *dc:Contributor* which represents a person or organization responsible for making these changes (See Fig. 2.5). Furthermore, each sub-layer is connected to the core layer (Sect. 2.3.2.1) through relations between their respective concepts. For instance, the concept *tei:Text* is equivalent to the concept *Text* of the core layer (See Fig. 2.6). *mpeg7:StillRegion* is sub-concept of *ImageElement* (See Fig. 2.8) and *dc:title* is sub-concept of *DescriptiveMetadata* (See Fig. 2.5). These relations allow concepts of existing standards (e.g., TEI, MPEG-7 and DC) to inherit common properties from the core layer.

2.3.2.3 Pluggable Domain-specific Layer

The previous layers model a heterogeneous document corpus independently of the content of the documents. We introduce the domain-specific layer as a pluggable layer to make the same version of *LinkedMDR* ontology easily adaptable to any domain-specific knowledge.

In the core layer, the concept *Domain* is linked to the concept *Object* through the relation *isRelatedTo* (See Fig. 2.9). This way, only two main steps are required at this stage in order to adjust *LinkedMDR* to a specific domain: (1) creating a sub-concept of *Domain* describing the application at hand (e.g., *Construction* for the AEC industry, *Medicine* for the medical domain), and (2) creating a sub-concept of the previously created concept (e.g., *IFC* as a sub-concept of *Construction*) that generalizes concepts of one or multiple external domain-specific ontologies. By means of inference rules, these concepts will be related to sub-concepts of *Object* (i.e., *Document*, *Media* and *MediaComponent*).

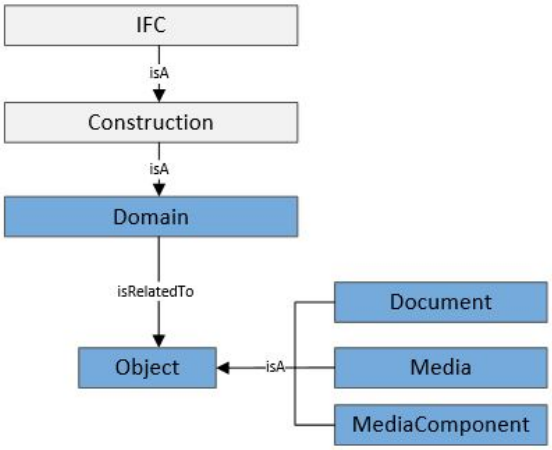


FIGURE 2.9 – Pluggability of a domain-specific layer in *LinkedMDR*.

Fig. 2.3 shows an example of the pluggable domain-specific layer where we integrated a simplified version of the ifcOWL [18] ontology. Fig. 2.10 shows another example of a domain-specific layer where we migrated into the medical domain and plugged in a portion of the Medical Subject Headings (MeSH) RDF schema[18].

A detailed example on *LinkedMDR* instances considering its different layers is explained in Sect. 2.3.3.2 and shown in Fig. 2.11.

2.3.3 Tightly Coupled Semantic Graph

In the literature, a tightly coupled semantic graph is known as a semantic graph where data objects (e.g., documents, web pages, tuples, etc.) and their metadata are individuals coupled with those of a lexical knowledge base or domain-specific ontology. Existing works [22, 81] have adopted tightly coupled semantic graphs in their

[18]MeSH provides a hierarchically-organized terminology for indexing and cataloging of biomedical information such as MEDLINE/PUBmed databases. It is available online at `ftp://nlmpubs.nlm.nih.gov/online/mesh/rdf/`

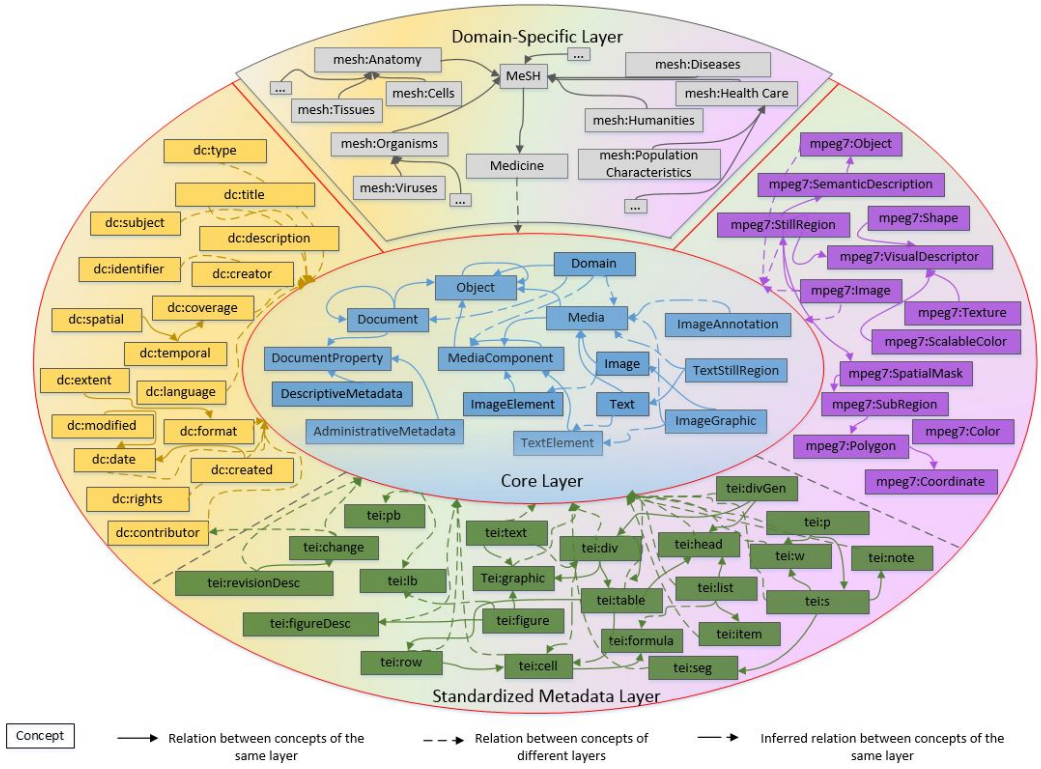


FIGURE 2.10 – Example of *LinkedMDR* application in the medical domain.

approaches, yet on different data types. For instance, Rocha et al. [81] coupled web pages and a domain-specific ontology. Chbeir et al. [22] coupled a lexical semantic network and a standard inverted index. Inspired by these approaches, we formally define our tightly coupled semantic graph which couples two external resources: a document corpus and a domain-specific ontology.

Sect. 2.3.3.1 formally defines the two underlying resources, Sect. 2.3.3.2 formally defines the resulting graph, and Sect. 2.3.3.3 describes the tight coupling algorithm that generates this graph based on the external resources, the infrastructure provided by *LinkedMDR* ontology and the associated middleware (indexer).

2.3.3.1 Underlying External Resources

We define a heterogeneous document corpus and a domain-specific ontology as follows:

Definition 1 (*Heterogeneous Document Corpus*)**.** A heterogeneous document corpus δ is defined as a set of n documents originated from different sources. The structure of the documents does not necessarily follow a common standard. Formally, $\delta = \{d_1, \ldots, d_n\}$. A document $d_i \in \delta$ is characterized by a set of p metadata (e.g., author, format, creation date, etc.) and a set of q media (e.g., text, image, video, or audio), such that $d_i = \{meta_1, \ldots, meta_p, med_1, \ldots, med_q\}$. Further, a media $med_l \in d_i$ is made of s media components (e.g., section, paragraph, etc. for the text; still region, object, etc. for the image), such that $med_l = \{medComp_1, \ldots, medComp_s\}$.

Definition 2 (*Domain-specific Ontology*)**.** Given an application domain $\mathcal{D}$, a domain-specific ontology designated as $\mathcal{O}_{\mathcal{D}}(C, R, Lit, A, L, f_L)$ is the semantic knowledge describing information in $\mathcal{D}$, where:

- C is the set of domain-specific concepts.

  For instance, considering a simplified version[19] of the ifcOWL [18] from the AEC industry. The ontology concepts include, but are not limited to, entities describing the building (*IfcBuilding*), the building elements (e.g., *IfcWindow*, *IfcDoor*, *IfcCurtainWall*) and the building systems (e.g., *IfcSystem* describing *Plumbing*, *Cooling*, *Fenestration*, etc.).

- $R \subseteq C \times C$ is the set of relations between domain-specific concepts in C.

  For instance, r_1 =*(IfcBuilding, IfcCurtainWall)* is a spatial containment relation between *IfcBuilding* and *IfcCurtainWall*, where $r_1 \in R$.

- $Lit = \{Integer, Decimal, String, ...\}$ is the set of literal types.

- $A \subseteq C \times Lit$ is the set of attributes describing domain-specific concepts i.e., relations between domain-specific concepts in C and literals in Lit.

  For instance, *WindowHeight* is an attribute property linking *IfcWindow* to *Decimal*.

- L is the set of relation labels.

- $f_L : R \rightarrow L$ is the association function that assigns a label $l \in L$ to a domain-specific relation $r \in R$, hence $f_L(r) = l$.

  For instance, $f_L(r_1) =$ *"contains"*.

For ease of presentation, $\mathcal{O}_{\mathcal{D}}(C, R, Lit, A, L, f_L)$ will be referred to as $\mathcal{O}_{\mathcal{D}}$ in the remainder of the thesis report.

2.3.3.2 Tightly Coupled Semantic Graph Model

We formally define a tightly coupled semantic graph as follows (Examples are based on Fig. 2.11 which resumes our motivating scenario depicted in Fig. 2.1):

Definition 3 (*Tightly Coupled Semantic Graph*)**.** Given a heterogeneous document corpus δ and a domain-specific ontology $\mathcal{O}_{\mathcal{D}}$, we define a *tightly coupled semantic graph* $\mathcal{G}_{\delta}(V, E, Val, f_{Val}, Lab, f_{Lab}, W, f_{W_v}, f_{W_e})$ as the instances graph describing the structural and domain-specific knowledge in δ following the infrastructure provided by *LinkedMDR*, where:

- V is the set of nodes representing instances of $\mathcal{O}_{\mathcal{D}}$ and δ:

  - $V = V_d \cup V_s$.

[19] A sample of the ifcOWL ontology concepts with simplified properties and relations.

- $V_d \subset V$ is the subset of domain-specific nodes where a node $v_d \in V_d$ represents an instance of $c \in C$ in $\mathcal{O}_\mathcal{D}$.

  For instance, *curtainwall4* is an instance of the concept *IfcCurtainWall*.

- $V_s \subset V$ is the subset of structural-based nodes where a node $v_s \in V_s$ represents any granularity element in δ i.e., a document $d_i \in \delta$ (e.g., *d4*), a metadata $meta_k \in d_i$ (e.g., *d4.title*), a media $med_l \in d_i$ (e.g., *d4.imagegraphic1* or more precisely a media component $medComp_r \in med_l$ (e.g., *d4.stillregion1*).

- $V_d \cap V_s = \emptyset$[20].

- E is the set of directed edges representing relations between nodes:

  - $E = E_d \cup E_s \cup E_h$.

  - $E_d \subseteq V_d \times V_d$ is the subset of domain-specific edges where an edge $e_d \in E_d$ represents an instance of a relation $r \in R$ in $\mathcal{O}_\mathcal{D}$.

    For instance, e_{d_1} = *(bldg1,curtainwall4)* is the spatial containment edge linking *bldg1* to *curtainwall4*.

  - $E_s \subseteq V_s \times V_s$ is the subset of structural-based edges where an edge $e_s \in E_s$ represents a relation between structural-based nodes in V_s (such as *part-whole*, *reference*, etc.), thus augmenting the representation of δ.

    For instance, e_{s_1} = *(d1, d1.div7)* is the edge linking *d1* to *d1.div7*.

  - $E_h \subseteq V_d \times V_s$ is the subset of hybrid edges where an edge $e_h \in E_h$ represents a tight coupling, i.e. a relation between a node $v_d \in V_d$ and a node $v_s \in V_s$.

    For instance, e_{h_1} = *(curtainwall4, d1.div7)* is the edge linking *curtainwall4* to *d1.div7*.

  - $E_d \cap E_s = \emptyset$, $E_s \cap E_h = \emptyset$, $E_d \cap E_h = \emptyset$, and $E_d \cap E_s \cap E_h = \emptyset$[21].

- *Val* is the set of node literal values.

  The value of a node $v_s \in V_s$ is made of its content.

  For instance, *"Exterior Facades"* is the string value associated to *d*1.*div*7.

  The value of a node $v_d \in V_d$ consists of its concatenated attributes $A_{v_d} \subseteq A$ in $\mathcal{O}_\mathcal{D}$ together with their associated values.

  For instance, *"label:facteur solaire; hasValue:0.64"* is the value associated to *solar-factor4* obtained from the concatenation of its attributes *label*, *hasValue* and their values *"solar factor"* and *0.64* respectively.

  For the sake of simplicity, we omit these values from the graph depicted in Fig. 2.11.

[20] We distinguish nodes of the subset V_s from those of the subset V_d to explicitly differentiate between structural characteristics and domain-specific ones.

[21] We also distinguish between edges of different subsets such as the case for nodes.

- $f_{Val} : V \rightarrow Val$ is the association function that assigns to a node $v \in V$ a literal value $val \in Val$, hence $f_{Val}(v) = val$.

  For instance, $f_{Val}(d1.div7) =$ *"Exterior Facades"*.

- Lab is the set of edge labels.

- $f_{Lab} : E \rightarrow Lab$ is the association function that assigns a label $lab \in Lab$ to an edge $e \in E$, hence $f_{Lab}(e) = lab$.

  For instance, $f_{Lab}(e_{s_1}) =$ *"hasPart"*.

- W is the set of both nodes and edges' weights.

- $f_{W_v} : V \rightarrow W$ is the node weight mapping that assigns a weight $w_v \in W$ to a node $v \in V$.

- $f_{W_e} : E \rightarrow W$ is the edge weight mapping that assigns a weight $w_e \in W$ to an edge $e \in E$. f_{W_e} consists of a set of functions which adapt to E_s, E_d, and E_h depending on different heuristics.

  The weight mapping functions f_{W_v} and f_{W_e} are used in the search process of query answers. They are detailed in Chapter 3, Sect. 3.4.4.

For ease of presentation, $\mathcal{G}_\delta(V, E, Val, f_{Val}, Lab, f_{Lab}, W, f_{W_v}, f_{W_e})$ will be referred to as $\mathcal{G}_\delta$ in the remainder of the thesis report.

Figure 2.11 shows an extract of a tightly coupled semantic graph representing the heterogeneous document corpus depicted in the motivating scenario of Sect. 2.1 (See Fig. 2.1), where each node and relation correspond respectively to instances of *LinkedMDR* concepts and relations of the three different layers. In the following, we demonstrate how this example of graph satisfies the challenges we mentioned in Sect. 2.1, regardless of the techniques used to generate it[22]:

- d_5 contains a figure which itself is an excerpt of the technical drawing d_4. This is represented by the following structural-based edges: e_{s_2} = *(d5, d5.div3.figure15)* with $f_{Lab}(e_{s_2})$ = *"hasPart"*, and e_{s_3} = *(d4,d5.div3.figure15)* with $f_{Lab}(e_{s_3})$ = *"includes"*. It represents an inter-document spatial relation between the two documents d_4 and d_5. This particularly addresses *Challenge 1.1*.

- d_4 contains several technical drawings, each related to a specific building floor and described on different pages of the document. As mentioned in Sect. 2.2, existing multimedia standard such as MPGE-7 [97] help in describing the different regions of the drawings but without any information on the corresponding pages. Existing text encoding standards such as TEI [98] can represent the pages of each drawing but without structural description of their content. Existing general metadata standards such as DC [29] can provide us with descriptors on d_4 as a whole entity without details on pages and content of each

[22] Details will be provided in Chapter 4.

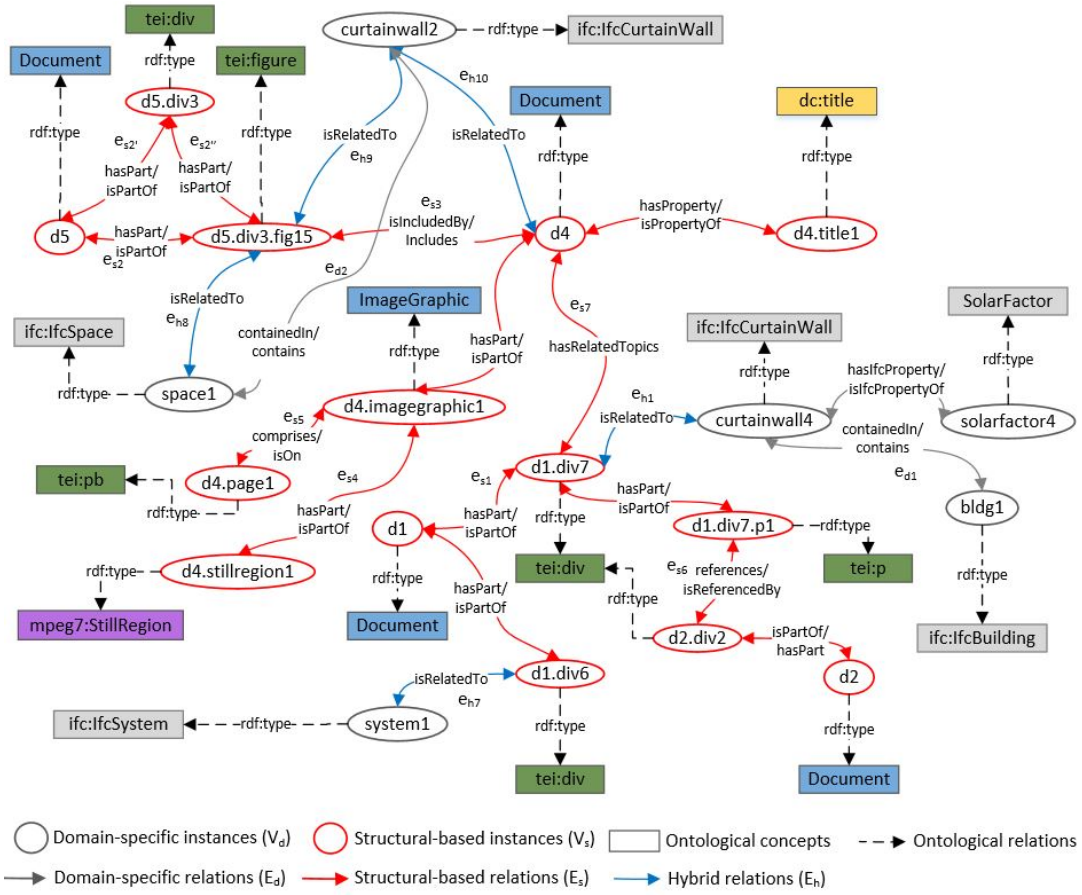


FIGURE 2.11 – Extract of a tightly coupled semantic graph representing the collection of heterogeneous documents in Fig. 2.1.

drawing. Through *LinkedMDR* core concepts and relations, it is possible to take advantage of the specialized features of existing standards while associating them to provide extended capabilities. For instance, this is represented by the structural-based node *d4.imagegraphic1* and its associated structural-based edges: e_{s_4} = *(d4.imagegraphic1,d4.stillregion1)* with $f_{Lab}(e_{s_4})$ = *"hasPart"*, and e_{s_5} = *(d4.imagegraphic1,d4.page1)* with $f_{Lab}(e_{s_5})$ = *"isOn"*. This particularly addresses *Challenge 1.2*.

- Section *6* of the document d_1, which describes the sanitary plumbing of the building, is represented by the structural-based node *d1.div6*. Section *7* of the document d_1, which describes the exterior facades, is represented by the structural-based node *d1.div7*. It is possible to associate these sections with their related domain-specific information. This is done through hybrid edges e_{h_1} = *(curtainwall4, d1.div7)* with $f_{Lab}(e_{h_1})$ = *"isRelatedTo"* and e_{h_7} = *(system1, d1.div6* with $f_{Lab}(e_{h_7})$ = *"isRelatedTo"* linking *d1.div7* and *d1.div6* to domain-specific nodes *curtainwall4* and *system1* representing respectively instances of concepts *IfcCurtainWall* and *IfcSystem* from the ifcOWL ontology [18]. Likewise in the document d_5, *d5.div3.figure15* represents a drawing related to the office space of a building floor. This is represented by e_{h_8} = *(space1, d5.div3.figure15)*

with $f_{Lab}(e_{h_8})$ = *"isRelatedTo"* and *space1* instance of *IfcSpace*. At this stage, the semantics associated to the different structural-based and domain-specific nodes, at different granularity levels, allows further inferred relations enriching the linked data graph. For instance, since the document d_4 contains the figure *15* of d_1, which is related to *space1*, then d_4 is also related to *space1*. Hence, a new hybrid edge e_{h_4} = *(space1, d4)* is created with $f_{Lab}(e_{h_4})$ = *"isRelatedTo"*. This particularly addresses *Challenge 1.3*.

- *d1*, *d4*, *d5* are instances of *Document* representing the technical specification Word document d_1, the technical CAD PDF document d_4 and the technical specification PDF document d_5 respectively. They have different media content types (text and image) and formats. This particularly addresses *Challenge 1.4*.

- All the nodes in the graph (See Fig. 2.11) are instances of *LinkedMDR* ontology: *d4* is an instance of the core concept *Document* of *LinkedMDR*; *d4.stillregion1* is an instance of the standardized metadata concept *mpeg7:StillRegion* of *LinkedMDR* which is provided by the MPEG-7 standard [97]; *curtainwall4* is an instance of the domain-specific concept *ifc:IfcCurtainWall* of *LinkedMDR* which is provided by the ifcOWL [18]. The nature of the ontology makes the graph easily extensible at different levels (new concepts, new relations, new instances). As an example, at the instances level, consider an external knowledge-base or ontology describing an audio file d_7 which represents the noise before and after the cladding of a given facade: *ext:facade*. If the latter finally corresponds to *curtainwall4*, then the two resources are linked through *owl:sameAs* relation. This particularly addresses *Challenge 1.5*.

2.3.3.3 Tight Coupling Algorithm

In this section, we present the pseudo code of our proposed tight coupling algorithm (See Algorithm 1). Although we give some examples on the techniques used by the middleware, we leave technical details regarding the underlying technologies, APIs and tools for Chapter 4. The algorithm takes as input: (i) a heterogeneous document corpus δ, (ii) a domain-specific ontology $\mathcal{O}_{\mathcal{D}}$, and (iii) *LinkedMDR* ontology. Its final output is the tight coupled semantic graph $\mathcal{G}_{\delta}$.

The overall process consists of six major steps:

- Step 1 (line 2): It uses existing automatic metadata extraction [42] and text engineering techniques [64] offered at the middleware indexing level. The former automatically generates metadata information of the document corpus δ (e.g., *title1* of d_4, *div7* and *div6* of d_1). The latter automatically generates some dependencies encountered in the text based on pre-defined regular expressions (e.g., the cross-reference dependency between *p1* of *div7* of d_1 and *div2* of d_2 encountered in the text of *p1*: *"performances based on section 2.2.11 in the thermal*

Algorithme 1 : Tight Coupling

Inputs : Heterogeneous document corpus δ; Domain-specific ontology $\mathcal{O}_\mathcal{D}$; *LinkedMDR* ontology.
Output : Tightly coupled semantic graph $\mathcal{G}_\delta$.

1 $\mathcal{G}_\delta \leftarrow \emptyset$; // initializes graph
// **STEP 1** using techniques for automatic metadata extraction and text engineering
2 $Output_s \leftarrow middleware(\delta)$; // generates structural-based descriptors of δ as output
// **STEP 2** using our tailored converters
3 $V_s \cup E_s \leftarrow middleware(Output_s, LinkedMDR)$; // converts standard output into ***LinkedMDR*** structural-based instances V_s and E_s relying on the semantics of the core and standardized metadata layers of ***LinkedMDR***
4 $\mathcal{G}_\delta \leftarrow \mathcal{G}_\delta \cup (V_s \cup E_s)$; // updates the graph with structural-based nodes and edges
// **STEP 3** using techniques for automatic semantic annotations
5 **foreach** $v_s \in V_s$ **do**
6 **if** $(f_{Val}(v_s)! = null)$ **then**
7 $Output_d \leftarrow middleware(f_{Val}(v_s), \mathcal{O}_\mathcal{D})$; // generates semantic annotations as output from the non null direct textual content of each generated v_s
 // **STEP 4** using our tailored converters
8 $V_d \cup E_d \leftarrow middleware(Output_d, LinkedMDR)$; // converts standard output into ***LinkedMDR*** domain-specific instances V_d and E_d relying on the semantics of the core and domain-specific layers of ***LinkedMDR***
9 $\mathcal{G}_\delta \leftarrow \mathcal{G}_\delta \cup (V_d \cup E_d)$; // updates the graph with domain-specific nodes and edges
10 **foreach** $v_d \in V_d$ **do**
 // **STEP 5** coupling domain-specific node to its related structural-based node
11 $e_h \leftarrow hybridCoupling(v_d, v_s, LinkedMDR)$; // creates hybrid edge between v_d and v_s using the semantics of the core layer of ***LinkedMDR***
12 $\mathcal{G}_\delta \leftarrow \mathcal{G}_\delta \cup \{e_h\}$; // updates the graph with previously created hybrid edge
13 **end**
14 **end**
15 **end**
// **STEP 6** running the reasoner
16 $\mathcal{G}_\delta \leftarrow runReasoner(\mathcal{G}_\delta, LinkedMDR)$; // dynamically creates inferred relations enriching $\mathcal{G}_\delta$ based on semantic rules in ***LinkedMDR***
17 **return** $\mathcal{G}_\delta$;

report"). This outputs standard metadata output ($Output_s$) mostly from existing standards for document's structure and content description (See Sect. 2.2), thus describing the structural part of the corpus δ.

- Step 2 (lines 3-4): It converts the generated descriptors into structural-based instances of *LinkedMDR* by using well-defined converters offered at the middleware indexing level. These converters are based on the semantics provided by the core layer and the standardized metadata layer of *LinkedMDR* i.e., their concepts, relations and semantic rules. This generates nodes V_s and E_s which are then added to the empty graph $\mathcal{G}_\delta$ (e.g., structural-based nodes *d1.div7.p1* and *d2.div2*, and edge e_{s_6} = *(d1.div7.p1, d2.div2)* with $f_{Lab}(e_{s_6})$ = *"references"*, See Fig. 2.12).

- Step 3 (lines 5-7): It looks over the direct textual content of each generated structural-based node $v_s \in V_s$ ($f_{Val}(v_s)$), as long as it is not empty, to identify one or multiple domain-specific instances $v_d \in V_d$ based on the knowledge embedded in a domain-specific ontology $\mathcal{O}_\mathcal{D}$. This is done using existing automatic techniques for semantic annotation [26] offered at the middleware indexing level. The rationale behind is to benefit from the automatic extraction of domain-specific information from textual content at different levels of

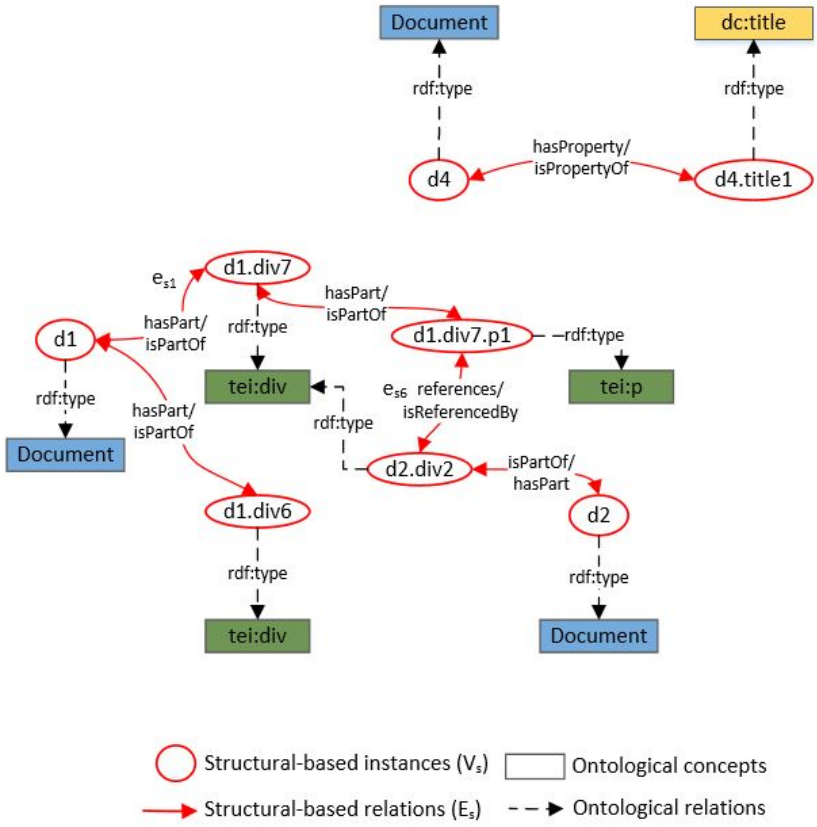


FIGURE 2.12 – Example of generated structural-based nodes an edges of the tightly coupled semantic graph depicted in Fig. 2.11 following Steps 1 and 2 of Algorithm 1.

δ, including the content of metadata descriptions, documents, sections, paragraphs, descriptions of images, etc. This generates annotations ($Output_d$) describing domain-specific information (e.g., *curtainwall4* identified as instance of *ifcCurtainWall* from the textual content *val* of *d1.div7*: $val = f_{Val}(d1.div7) =$*"exterior facades"*.

- Step 4 (lines 8-9): It converts the generated descriptors of the previous step into domain-specific instances of *LinkedMDR* by using well-defined converters offered at the middleware indexing level. These converters are based on the semantics provided by the core layer and the domain-specific layer of *LinkedMDR* i.e., their concepts, relations and semantic rules. This generates nodes V_d and E_d which are then added to the graph $\mathcal{G}_\delta$. Note that, information on the textual term or sequence of terms that allowed the identification of a domain-specific concept is added as an attribute *label* of the domain-specific instance, thus it is part of *val*, where $val = f_{Val}(v_d)$ as explained in Definition 3 (e.g., $f_{Val}(curtainwall4) =$*"label:exterior facades"*, See Fig. 2.13). Also note that, if two domain-specific instances in V_d are identified in the same structural-based instance v_s and there exists in *LinkedMDR*'s domain-specific layer one or several relations between them, then these relations are automatically created between them as domain-specific edges in E_d (e.g., $e_{d_2} = (curtainwall2, space1)$ with

$f_{Lab}(e_{d_2})$ = *"isContainedIn"* is created since *curtainwall2* and *space1* are both identified in the textual description associated to the same structural-based node *d5.div3.fig15*).

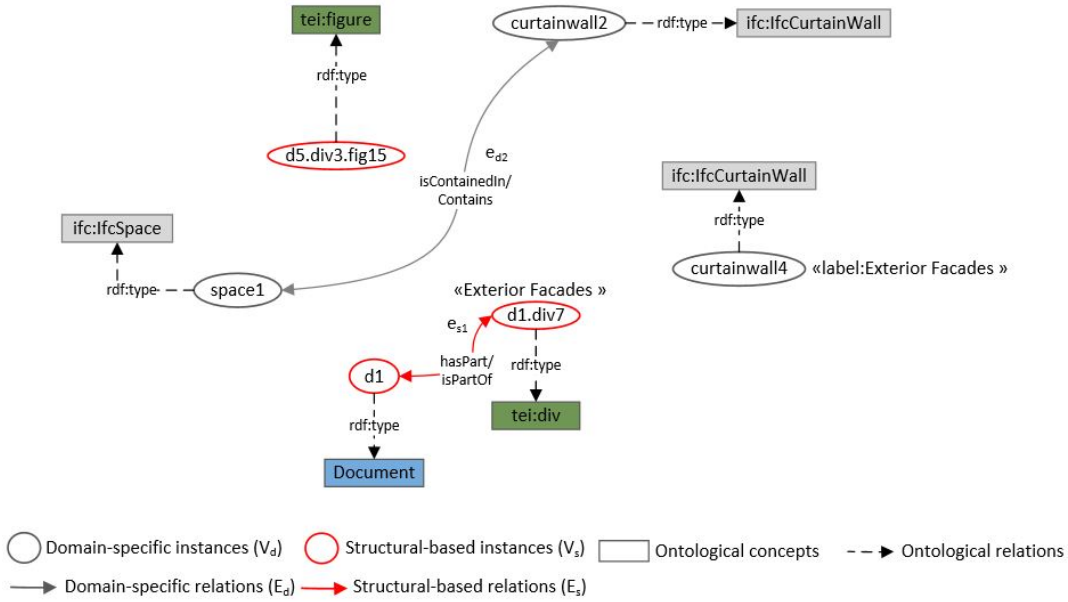


FIGURE 2.13 – Example of generated domain-specific nodes and edges of the tightly coupled semantic graph depicted in Fig. 2.11 following Steps 3 and 4 of Algorithm 1.

- Step 5 (lines 11-12): It creates the hybrid coupling i.e., the hybrid edge e_h linking domain-specific node v_d to its corresponding structural-based node v_s where it was identified in the previous step. The rationale behind is that if a node v_d where identified from the direct textual content of a node v_s then it is directly related to it as it provides domain-specific information describing its content. The linking process is based on the semantics provided by the core layer of *LinkedMDR* precisely by the relation *isRelatedTo* between a concept *Domain* and *Object* (See Sect. 2.3.2.1). The generated hybrid edge e_h is finally added to the graph $\mathcal{G}_\delta$ (e.g., e_{h_1} = *(d1.div7, curtainwall4)* with $f_{Lab}(e_{h_1})$ = *"isRelatedTo"*, See Fig. 2.14).

- Step 6 (line 16): It runs the semantic reasoner so that further inferred relations are dynamically added to the graph $\mathcal{G}_\delta$ based on the semantic rules provided by *LinkedMDR*. For instance, the relation *hasPart* is transitive, thus given the two edges $e_{s'_2}$ = *(d5, d5.div3)* with $f_{Lab}(e_{s'_2})$ =*"hasPart"* and $e_{s''_2}$ = *(d5.div3, d5.div3.fig15)* with $f_{Lab}(e_{s''_2})$ =*"hasPart"*, the edge e_{s_2} = *(d5, d5.div3.fig15)* with $f_{Lab}(e_{s_2})$ =*"hasPart"* is dynamically inferred after the reasoner is started (See Fig. 2.15). Furthermore, given the two edges e_{h_9} = *(curtainwall2, d5.div3.fig15)*

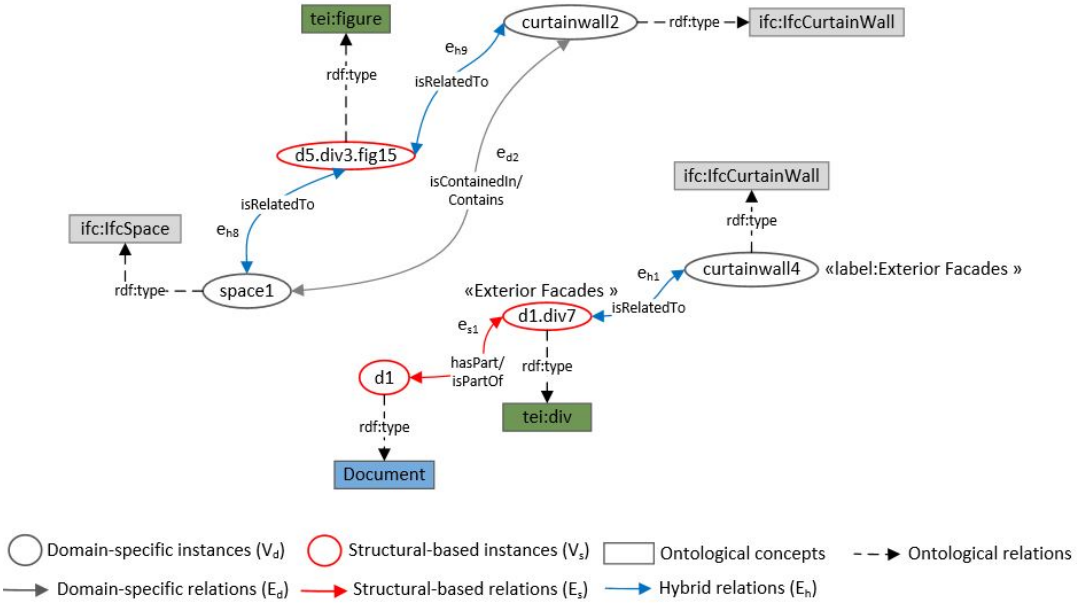


FIGURE 2.14 – Example of a generated hybrid edge of the tightly coupled semantic graph depicted in Fig. 2.11 following Step 5 of Algorithm 1.

with $f_{Lab}(e_{h_9}) =$ *"isRelatedTo"* and $e_{s_3} =$ *(d4, d5.div3.fig15)* with $f_{Lab}(e_{s_3}) =$ *"includes"*, the edge $e_{h_{10}} =$ *(curtainwall2, d4)* with $f_{Lab}(e_{h_{10}}) =$ *"isRelatedTo"* is dynamically inferred (See Fig. 2.15). The latter, together with the edge $e_{h_1} =$ *(curtainwall4, d1.div7)* with $f_{Lab}(e_{h_1}) =$ *"isRelatedTo"*, also infer the edge $e_{s_7} =$ *(d4, d1.div7)* with $f_{Lab}(e_{s_7}) =$ *"hasRelatedTopics"* since *curtainwall1* and *curtainwall2* are both instances of the same Concept i.e., *IfcCurtainWall* (See Fig. 2.15).

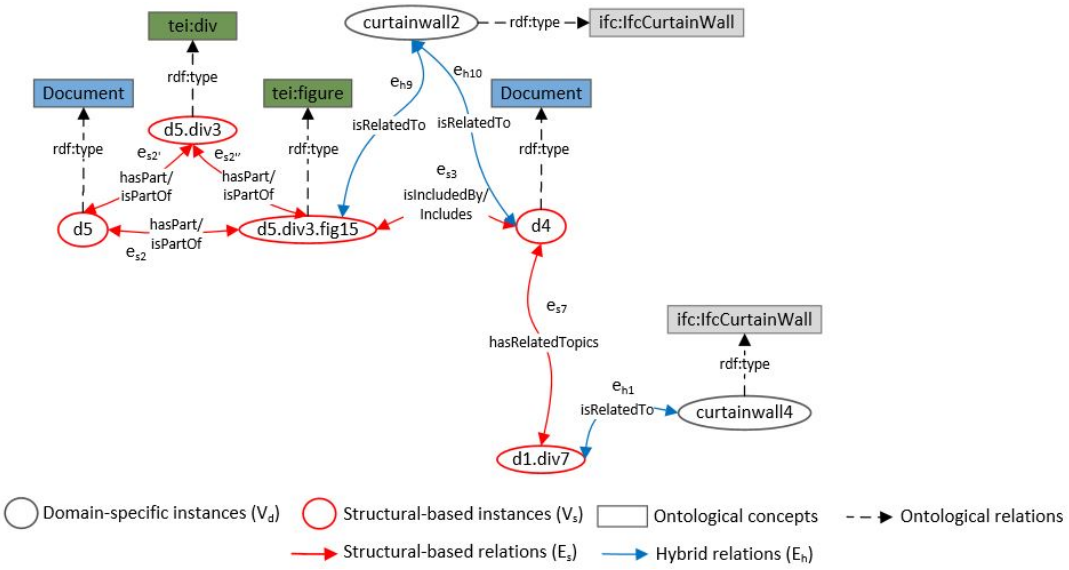


FIGURE 2.15 – Example of generated inferred edges of the tightly coupled semantic graph depicted in Fig. 2.11 following Step 6 of Algorithm 1.

2.4 Summary

This chapter tackles the problem of representing a heterogeneous document corpus for it to be exploited in a SIR system, which has been partially solved by existing standards and data models for document representation. We introduce a novel semantic-based approach, which we call *Tight Coupling* approach, aiming to represent the collective knowledge embedded in a given heterogeneous document corpus through a *tightly coupled semantic graph*. The contributions of this chapter come down to: (i) defining a backbone multi-layered ontology, entitled *LinkedMDR* [20], which provides the required infrastructure to build this graph through core components easily connected to existing standardized metadata standards and easily pluggable to domain-specific ontologies, (ii) defining the tightly coupled semantic graph that embeds instances of *LinkedMDR* where semantics are associated to structural components of the documents as well as their domain-specific information, and (iii) defining the *tight coupling algorithm* which generates this graph. To our knowledge, this supports the first graph capable of: (1) representing the various inter and intra-document links within the corpus, (ii) describing general metadata information of the document, its content, and its structural text and multimedia components all combined, (iii) associating semantics at content and structural levels of the document, (iv) handling document multimodality, and (v) ensuring extensibility. Also, the proposed approach is modular and generic, so it does cope with any domain-specific application. Several experiments were conducted to validate our proposal within real-world applications. These experiments are dedicated to Chapter 4.

Chapter 3

Information Retrieval over a Heterogeneous Document Corpus

"Real search is about providing valuable information when it's really needed to those who are actually looking for it."
- David Amerland

The increased availability of interdependent heterogeneous data generated from different sources is fostering the incorporation of semantic knowledge-based graphs in information management and search applications. One of the utmost challenge remains in searching for relevant information among heterogeneous interdependent documents. One of the major limitations of the existing SIR systems is that they mainly rely on semantic graphs representing lexical and domain-specific information contained in heterogeneous data, without considering the structural-based components of the documents and dependencies between them. We consider that relying on tightly coupled semantic graphs is one step forward towards overcoming this major problem as they handle the representing of such information from the design phase. **This Chapter tackles the problem of searching for relevant information from tightly coupled semantic graphs while augmenting the search results with meaningful context including both structural and domain-specific dimensions of a heterogeneous document corpus.** We propose a novel data structure for query answers based on well-defined sub-graphs, which we call *Hybrid Molecules*, extracted from a tightly coupled semantic graph. We also provide a comprehensive query processing pipeline, entitled *HM Query Processing*, based on the defined hybrid molecules. Its main purpose is to generate a ranked list of the desired hybrid molecule-based query answers based on the user's query. Our main contributions, within the hybrid molecule-based query processing, are located in the Search and Ranking modules and they consist of: *HM_CSA*, a novel graph-based search algorithm over a tightly coupled semantic graph, and *Weight Mapping* functions which score the components of the hybrid molecules and rank the query answers conveniently.

3.1 Introduction

Web and Information Systems are increasingly adopting semantic knowledge-based models to represent the data encapsulated in heterogeneous resources [7, 49]. This has several proven benefits in improving users' experience in search applications [62]. As mentioned in Chapter 1 and Chapter 2, in several industries involving multidisciplinary projects, users looking for a particular information have to search through interdependent heterogeneous documents provided by different sources. Fig. 3.1 illustrates a sample of two interdependent documents from the AEC industry recalling an extract of the motivating scenario presented in Sect. 2.1 of Chapter 2: d_1 (a technical specification document) and d_2 (a material thermal report) related to the same project. Parts of document d_1 describe acoustic and thermal properties (*sub-section 1* and *sub-section 2* of *section 7* respectively). Document *d2* describes a thermal study which includes details on the solar factors of windows (*table 1* of *section 2*), implicitly described by *"SW coefficient*[1]*"*. The reference relation between the two documents shows that *section 2* of document d_2 contains information complementary to *sub-section 2* of document d_1.

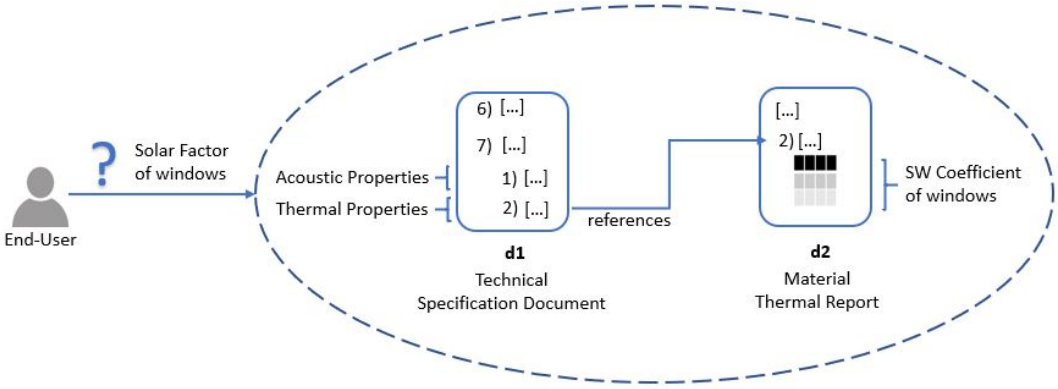


FIGURE 3.1 – Extract of Fig. 2.1: sample documents from the AEC industry.

Consider that the user is searching for *"the solar factor of windows"*. Several challenges arise in order to provide him with relevant query answers, that do not only include documents, but also more refined and enriched information (See Fig. 3.2):

- *Challenge 2.1: Providing relevant granularity levels of the documents* - Relevant granularity levels of the documents associated to domain-specific information searched by the user (e.g., *section 2* of document *d*2 entitled *"2) Characteristics of Window Frames"* describing the value of the solar factor i.e., "$SW = 0.45$") help the user in locating desired information from possibly large documents involving other irrelevant sections.

- *Challenge 2.2: Providing relevant inter and intra-document dependencies* - Relevant dependencies between documents and parts of documents (e.g., *sub-section 2* of

[1]SW stands for Short Wavelength shading coefficient and represents the capacity of the glazing to transmit the solar heat inside the building.

document d_1 entitled *"Thermal Properties"* that involves additional information regarding the solar factor coefficients described in *section 2* of document d_2) enrich the search results as they provide further related and useful information across the documents.

- *Challenge 2.3: Presenting contextualized query answers* - Contextual information related to the relevant information covers both dimensions of the documents i.e., the structural-based dimension (e.g., *"Page 3"* of document d_2) and the domain-specific dimension (e.g., the heat transfer coefficient) explained in Chapter 2. The query answer and its contextual information should be presented in a meaningful structure to help the user in interpreting the results and tracking cross-document dependencies, which reduces their efforts, wasted time and errors.

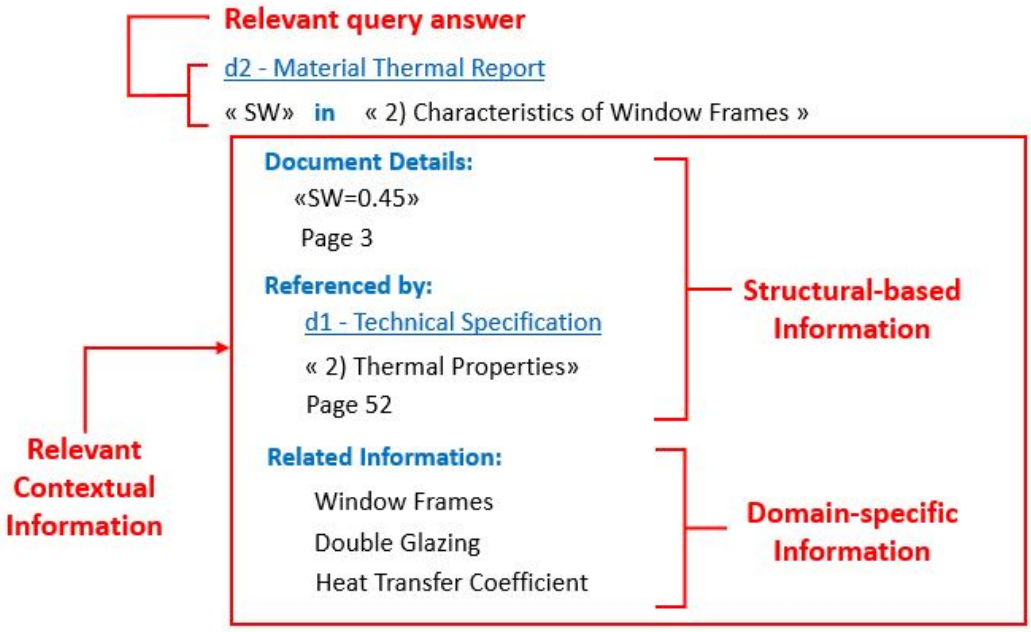


FIGURE 3.2 – Example of a contextualized query answer regardless of the display methods.

Traditional IR approaches mainly rely on syntactic keyword-based search [63]. To overcome their limitations, there has been significant interest in taking semantics into account leading to the emergence of SIR systems. Although suitable for several applications [23, 32, 38, 56, 81, 95, 108], SIR systems provide documents as query answers without considering in their search results (i) detailed information regarding relevant granularity levels of the documents, (ii) various inter and intra-document dependencies, and (iii) relevant contextual information. To the best of our knowledge, none of the existing approaches have tackled the above challenges combined.

In this chapter, we provide a solution to the aforementioned limitations of current SIR systems. As we already discussed, in Chapter 2, the importance of adopting a tightly coupled semantic graph to represent the collective knowledge embedded

in a heterogeneous document corpus, we take advantage of such a model to provide a novel data structure for query answers, which we call *Hybrid Molecules*. The latter consist of hybrid sub-graphs encapsulating domain-specific information coupled with related structural-based information of the documents. The hybrid molecule-based query answers bring in helpful contextual information of the documents improving the search results and reducing users' efforts in tracking and interpreting them. We formally define the hybrid molecule's structure in view of the characteristics of a tightly coupled semantic graph and the definition of a molecule concept in the literature [27, 28, 30, 36, 69]. We then integrate the notion of hybrid molecules in a query processing pipeline, where users submit their natural language (e.g., plain English text) queries over a heterogeneous document corpus and obtain relevant answers in the form of hybrid molecules. Although we present a bunch of hybrid molecule-based algorithms for each stage of the query processing, we focus on the graph-based search algorithm which generates a list of hybrid molecules, and the weight mapping which introduces weighting functions to rank the molecules conveniently.

The contributions of this chapter can be summarized by:

- *Hybrid molecule*, a novel data structure for query answers based on tightly coupled semantic graphs;
- *HM Query processing*, a comprehensive query processing pipeline over a heterogeneous document corpus, where the *hybrid molecule* structure intervenes in each of its stage;
- *HM_CSA*, an algorithm that constructs relevant hybrid molecules from a tightly coupled semantic graph;
- *Weight mapping*, a series of weighting functions that rank the hybrid-molecule query answers.

The remainder of this chapter is structured as follows. Sect. 3.2 provides fundamental notions of IR and reviews the related work regarding traditional IR and SIR systems. In Sect. 3.3, we introduce the *Hybrid Molecules* and formally define them based on a tightly coupled semantic graph's definition. Sect. 3.4 details on the *hybrid molecule-based query processing* pipeline over a heterogeneous document corpus with a particular focus on the underlying *HM_CSA* algorithm and *Weight Mapping* for search and ranking modules respectively. Sect. 3.5 concludes the chapter.

3.2 Background and Related Work

In this section we first present some basic background concepts in IR and the classical pipeline behind (Sect. 3.2.1). Then, we provide a literature overview on the different existing IR models and the existing approaches which implement them. We mainly distinguish between 2 categories of IR models: traditional IR models

(Sect. 3.2.2) and SIR models (Sect. 3.2.3). We then focus on the approaches based on SIR models since we rely on semantic graphs and ontologies as mentioned in Chapter 2.

3.2.1 Information Retrieval (IR)

IR is a wide area in Information Science and has been subject to many research works for decades. The main purpose of an IR system is to search and retrieve relevant resources from a collection of resources in order to satisfy user needs expressed as queries. IR systems have significantly evolved especially with the emergence of web search engines [1]. However, IR systems mainly rely on the following classical pipeline[2] for query processing (See Fig. 3.3):

- *Query Interpretation:* the system understands the user's query expressed in terms of keywords, natural language text or visual components and translates it into its internal knowledge structure. This sometimes includes preprocessing techniques (such as NLP for natural language text) and query modification (such as query re-writing) to increase both precision and recall [62].
- *Search:* the system searches for relevant information that matches the user's query from the descriptors (e.g., inverted lists, graphs, etc.) representing the collection of resources. The system retrieves partially or exactly matched resources depending on the underlying adopted IR model.
- *Ranking:* the retrieved resources are given scores according to their degree of relevance w.r.t. the user's query. The system ranks the results based on these scores.
- *Presentation:* the system presents the results to the user in response to their original query, within a Graphical User Interface (GUI), in an understandable format (such as SERP).

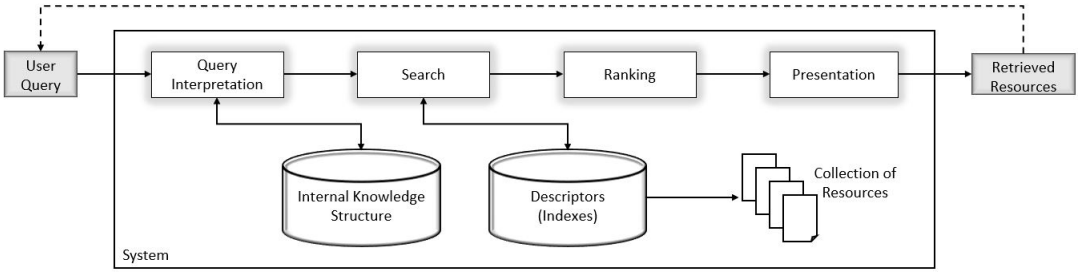


FIGURE 3.3 – Classical pipeline for query processing in IR.

[2]Inspired by the general model of IR presented by Belkin and Croft [8].

3.2.2 Traditional IR models

The classical IR approaches are based on Boolean, Vector Space and Probabilistic models [8, 63]. They mainly focus on the bag of words theory to represent the documents without considering the context of these words nor the different meanings that might be associated to them [55]. In the following, we first describe these models, then we provide examples of existing works and systems built on them.

The Boolean retrieval is based on the exact match between the set of terms representing the documents and the Boolean expressions[3] representing the query [48]. In this model, there is no distinction between the retrieved documents i.e., there is no form of relevance ranking. The Vector Space and Probabilistic retrieval are both based on the best match theory. The Vector Space Model (VSM) [86] is the most widely used retrieval model. It consists of representing the documents and the query by means of vectors of weighted terms, where weights are calculated using statistical distributions such as the term frequency-inverse document frequency (tf-idf) function. Document vectors are then compared to the query vector and assigned scores representing the relevance of documents to the query. The Probabilistic retrieval model [35] is based on the Probability Ranking Principle (PRP). The documents are ranked based on the probability of relevance of their texts to the query. This probability is often calculated using the Bayes Theorem[4].

Despite the drawbacks of the Boolean retrieval model, the latter was the main adopted retrieval model for decades until the arrival of the WWW. At this point, extended models have started to emerge. For instance, Salton et al. [85] propose an extended Boolean model to overcome the drawbacks of the standard Boolean model and improve the effectiveness of the search results. They combine the characteristics of VSM with the properties of Boolean algebra and calculate the similarity between queries and documents, so as to cover partial matching and ranking. Turtle and Croft [100] are the first to introduce the use of Bayesian networks in IR to represent probabilistic dependencies between a document and a user query by means of a directed and sophisticated graph. They combine Boolean, statistical and probabilistic models. Haines and Croft [47] extend this work by including relevance feedback techniques. This aims to improve retrieval performance for a particular query by modifying the query based on the user's reaction to the initial retrieved documents. In other words, the user judges the relevance or non-relevance of some of the documents retrieved and this feedback is later used to add new terms to the query and to re-weight query terms.

On the other side, there exist many open-source systems, toolkits and platforms on the basis of the aforementioned IR models. Lucene[5] is an IR system which uses a combination of the Boolean model and VSM. It essentially remains a VSM-based

[3] Words or phrases combined using the standards operators of Boolean logic.

[4] The Bayes Theorem describes the probability of a term appearing in a document, based on prior knowledge of conditions that might be related to the document such as its relevance.

[5] `http://lucene.apache.org/`

system, but uses the Boolean model to first narrow down the documents that need to be scored based on the use of Boolean logic in the Query specification [39]. It also adds some capabilities and refinements onto this model to support Boolean and fuzzy searching. Lemur[6] is an IR toolkit that is based on VSM and Probabilistic IR models. It offers several retrieval algorithms including simple tf-idf, Okapi (BM25) weighting scheme, and Kullback-Leibler (KL) divergence measure [63]. Terrier[7] is a modular platform for the rapid development of large-scale IR applications. It supports several IR approaches such as Divergence From Randomness, BM25F, and dependence proximity. It also offers supervised ranking models via Learning to Rank techniques [72].

3.2.3 SIR models

Many research works have attested the benefits of incorporating knowledge bases and domain specific ontologies in their index, query, and search process to overcome the semantic gap between keywords found in the documents and those in the user's query [38]. The main advantage of such methods resides in maximizing the precision and recall w.r.t. to traditional IR [63] where the search is limited to syntactic keyword matching. In the literature, there have been many attempts to provide classification categories for SIR models [62, 78]. However, this depends on various and numerous criteria such as the annotation models behind, the semantic similarity measures used to compare concepts representing the queries with those representing the documents (e.g., Resnik [80], Wu and Palmer [106]). In general, independently of the classification criteria used, SIR models rely on the conceptual search which replaces the bag of words paradigm with the bag of concepts outstripping traditional IR models [19].

In this study, we focus on existing works which are based on SIR models. We distinguish between fully-fledged SIR approaches which consider semantics in their index, query and search approaches (See Sect. 3.2.3.1) and approaches tackling specific issues in SIR, more precisely (i) dealing with the complexity of expressing SPARQL queries for users with limited technological skills (Sect. 3.2.3.2), (ii) relying on graph-based strategies to search for relevant information (Sect. 3.2.3.3), and (iii) integrating the notion of molecules to represent particular sub-graphs structure (Sect. 3.2.3.4).

3.2.3.1 Fully-fledged approaches

The first attempt towards injecting semantics in IR was the adoption of Word Sense Disambiguation (WSD) techniques [67] where terms are associated to concepts of a thesaurus or knowledge base such as WordNet[8]. Mihalcea and Moldovan [66]

[6] `https://www.lemurproject.org/`

[7] `http://terrier.org/`

[8] An electronic lexical database where words are grouped into sets of synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

extend the Boolean IR model by adding word semantics to free-text index. Their approach expands both the queries and the indices with WordNet synsets provided by a semi-complete yet highly precise WSD algorithm. The disambiguation process is based on pairs formed by the word to disambiguate and neighboring words in its context and their co-occurrence in SemCor, a semantic tagged corpus. Tekli et al. [95] build upon the idea of semantic aware search to target textual databases. They propose a general framework for modeling and processing semantic-aware querying. They first generate a semantic-aware inverted index structure, called SemIndex. It consists of a tightly coupled inverted index graph that combines a semantic network (such as WordNet) and a standard inverted index on a collection of textual data. They also provide a general keyword query model and query processing algorithms to retrieve semantic-aware results.

Kiryakov et al. [56] introduce a semantic platform, called the Knowledge and Information Management (KIM), for information extraction and retrieval over large document collections. Their semantic annotation is based on the Named Entity (NE) recognition i.e., the process that assigns to the entities encountered in the text (web pages, non-web documents, text fields in databases, etc.) links to their semantic descriptions. This is done based on an underlying ontology, called KIM Ontology (KIMO). Documents are then retrieved on the basis of relevance to NEs instead of words. Chechev et al. [23] present a prototype for patent retrieval within the Molto project. It addresses high quality domain-specific machine translation on web documents covering up several languages. It is based on a semantic tagger that annotates the patent content using a patent-specific ontology. In the search results, the tool provides a list of classes from the ontology that match the query and a list of matching documents. It also offers links to access the semantically annotated documents and the original patents. In the text of annotated documents, there are highlights on the words that are related to any semantic item. Fernandez et al. [32] present an ontology-based IR model that takes advantage of domain knowledge bases to support semantic search capabilities over large document repositories such as the web environment. The proposed system takes a formal SPARQL query as input, then executes the query against a knowledge base. This returns a list of matching semantic entities that satisfy the query based on the exact match strategy. Finally, the documents that are annotated with the previously returned instances are retrieved, ranked, and presented to the user based on an approximate match strategy.

3.2.3.2 SPARQL-based approaches

RDF is the widely used data model for representing semantic graphs [104]. Simple Protocol And Rdf Query Language (SPARQL) [105] is the predominant language used to query and manipulate RDF graph content on the Web or in an RDF store. Writing SPARQL queries is a tedious task and requires a technical knowledge in RDF and SPARQL syntax. It also requires full knowledge of the complex and heterogeneous structure of the linked data and handling its evolution over time. In

order to enable users with limited technological skills to express their information retrieval requirements and query linked data, several works have been defined to support visual querying, query rewriting and refinement.

Haag et al. [46] propose a visual querying framework for the formulation of SELECT and ASK SPARQL queries without any text input. In another work [45], they develop QueryVOWL, a visual query language which defines SPARQL mappings for graphical elements of the ontology visualization VOWL. Benedetti et al. [10] present LODeX, an interactive tool that provides the users with a summary view of the dataset structure and supports them in creating a visual query and refining it. Soylu et al. [90] present an ontology-based visual query system, called OptiqueVQS, that helps domain experts users in formulating visual queries for industrial applications.

Frosini et al. [34] propose a flexible query processing approach. It is based on query rewriting using query approximation and relaxation operators. These operators are defined within SPARQL^{AR}, an extension of a fragment of SPARQL 1.1[9]. Query Approximation consists of applying edit operators (such as deletion, insertion and substitution) to transform a well-defined regular expression into a new one. Query Relaxation relies on a fragment of RDF Schema (RDFS) entailment rules. In general, Query Relaxation is one of the most used cooperative techniques that provide users with alternative answers instead of an empty result. Fokou et al. [33] approach this problem by finding subqueries responsible of the failure, called Minimal Failing Subqueries (MFSs). These subqueries explain the empty returned result and guide the user to perform the relaxation process. They also compute a particular type of RDF queries, called Maximal Succeeding Subqueries (XSSs), which are subqueries with maximal number of triple patterns of the initial query. The rationale behind is that the set of triple patterns that are not in an XSS could be removed or made optional in the relaxed query. Angles et al. [3] propose an extension of SPARQL, called SPARQL$_{Ext}$ which preserves the original semantics while incorporating all known types of subqueries in a modular fashion: subqueries in FROM clauses, subqueries as graph patterns, and subqueries in filter constraints (Set membership, quantified and existential conditions).

3.2.3.3 Graph-based approaches

As previously mentioned, SIR systems rely on semantic graphs representing the data at hand. Thus, the retrieval process mainly consists of finding the relevant information from the graph substituting the data. Existing works have approached this problem differently. Some of them adopted subgraph matching approaches which consist of representing the data and the query through graphs G and Q respectively and then retrieve all sub-graphs of G that are similar to Q. Other approaches adopted graph exploration strategies where query terms are used to find relevant

[9]The latest version of SPARQL, available at `https://www.w3.org/TR/sparql11-overview/`

nodes in the graph. From these nodes, an algorithm traverses the graph to determine semantically related resources.

Zoo et al. [109] represent SPARQL queries as query graphs. They answer the query following a subgraph matching approach. The matching is based on finding, in an RDF graph, sets of connected vertices that are similar to the one representing the query. Their proposal consists of a graph-based SPARQL query engine, called gStore. Its main purpose is to handle, in a uniform and scalable manner, SPARQL queries with wildcards[10] and aggregate operators over dynamic RDF datasets. Zhong et al. [108] propose an approach for semantic search by matching pairs of Conceptual Graphs (CGs): the first one representing the query graph and the second one representing the candidate resource graph from web pages describing garment shops. A query CG has an entry representing a central word pointed by the user from the query. Candidate resource CGs are those which entries (concepts representing garments) are mapped to the query CG entry. Hierarchical relations in WordNet are considered for this mapping. The matching algorithm is based on both semantic similarities between central entries of two CGs and semantic similarities between sub-graph pairs of CGs representing relations affiliated to these entries.

Among the many heuristic graph-based search methods [82], Constrained Spread Activation (CSA) has been widely adopted in many IR applications where it proved its effectiveness [62, 81]. It is based on a Breadth-First Search (BFS) graph traversal strategy. It works by spreading out the activation from a set of start nodes to adjacent nodes progressively in the graph until predefined constraints are met. Cohen and Kjeldsen [24] use CSA algorithm to realize intelligent matches between user requirements and relevant agents in a Q&A application. Crestani et al. [25] were the first to apply CSA technique to the World Wide Web to retrieve information using an ostensive approach to querying similar to query-by-example. Griffith et al. [43] propose to enhance users recommendations using a collaborative filtering approach. They rely on graph-based representation of the problem domain and a CSA approach. Sun et al. [93] combine CSA algorithm with a spatial ontology to improve results in associative retrieval of spatial big data. Gouws et al. [40] compute semantic relatedness by applying spread activation over the hyperlink structure of Wikipedia. One of the most prominent uses of CSA is the one proposed by Rocha et al. [81]. It consists of the combination of CSA techniques with classical search techniques for searching in the semantic graph of a given domain. The semantic graph consists of a tight coupling between web pages and concepts of a domain-specific ontology. Their query execution follows a hybrid approach: given a user query, a classical search engine identifies a set of matched nodes; these nodes are then used as the start nodes of a spreading activation algorithm which subsequently activates possibly relevant neighbors in the graph. At the end of the algorithm, nodes (i.e., documents) with highest activation values are ranked the highest in the result set.

[10] Flexible criteria added to the query.

3.2.3.4 Molecule-based approaches

Molecules, mainly RDF molecules, have received a wide attention over the past two decades within different applications.

It is first introduced by Ding et al. [28] for the purpose of tracking RDF provenance and evaluating trustworthiness against RDF data in SW applications. They define it as the finest and lossless connected subgraph decomposition of the original RDF graph based on Functional Properties (FP) and Inverse Functional Properties (IFP), which are specified in a background ontology. Della et al. [27] expand on the same concept of RDF molecules yet for Stream Reasoning applications. The main purpose is to reason, in real time, over ontological knowledge by combining data streams and reasoning techniques. They introduce RDF Molecule Streams to abstract an aggregation of sampled input data streams. An RDF Molecule Stream has a timestamp that denotes the end of the aggregation interval and the logical arrival time of the molecule on the outgoing stream for reasoning. Newman et al. [69] also use an extended version of the original definition of RDF molecules [28] for distributed querying using MapReduce. They propose a MapReduce-based RDF molecule store that decomposes the graph into molecules, uses SPARQL to query the molecules, and merges them to generate search results. In their definition of RDF molecules, they take into account hierarchy and ordering to overcome existing limitations in the decomposition and merging. For instance, the hierarchy feature helps in distinguishing single level triples with two blank nodes. The ordering feature allows for the rapid retrieval of triples.

Endris et al. [30] introduce RDF Molecule Templates (RDF-MTs) in the context of federated SPARQL query processing over RDF datasets. RDF-MTs are used to describe the structure of RDF datasets. They are then used to bridge between parts of a query to be executed in a federated manner, thus guiding the source selection, query decomposition and optimization. The authors define an RDF-MT as an abstract description of entities belonging to the same RDF Class together with related properties and object properties. Consequently, internal and external links are created between RDF-MTs representing object properties from internal and external RDF datasets respectively. The proposed RDF-MTs improve the performance of the federation process. Galkin et al. [36] rely on the same definition of RDF molecules, however, the intention is to identify semantically equivalent RDF subgraphs. They present SJoin, a semantic similarity join operator to solve the problem of matching semantically equivalent RDF molecules from RDF graphs.

3.2.4 Discussion

Both existing traditional IR and SIR approaches use the same classical pipeline for query processing, yet they rely on different models and techniques as shown in Tables 3.1 and 3.2. This proves that the underlying 4 stages of this pipeline are the basis of any IR system.

TABLE 3.1 – Models and techniques adopted by existing traditional IR approaches w.r.t. the classical pipeline in query processing.

IR	Models & Techniques
Query Interpretation	Keywords, Boolean expressions, Vectors of terms NLP techniques, Query Modification
Search	Exact Match (Boolean Retrieval), Best Match (VSM, Probabilistic, Combined Retrieval)
Ranking	No ranking (for Boolean Retrieval), Scores of vectors/terms (e.g., tf-idf, BM25)
Presentation	Resources matching the query (for Boolean Retrieval), Ranked list of resources relevant to the query

TABLE 3.2 – Models and techniques adopted by SIR approaches w.r.t. the classical pipeline in query processing.

SIR	Models & Techniques
Query Interpretation	Keywords, Concepts, Graphs Visual SPARQL, NLP, Semantic Annotation techniques, Query Modification (Disambiguation, Relaxation, etc.)
Search	Best Match (VSM, Probabilistic, Combined Retrieval) Graph traversal techniques (e.g., CSA), Sub-graph matching, etc.
Ranking	Scores of vectors/terms/concepts/nodes/graphs (e.g., using semantic similarity measures)
Presentation	Ranked list of resources relevant to the query

To recap, we summarize the adopted models and techniques w.r.t. each stage as follows:

- *Query Interpretation:* Traditional IR approaches basically use keywords, Boolean expressions and vector of terms to express user's queries [63]. A major breakthrough has been achieved with SIR approaches using knowledge-based concepts and semantic graphs to represent these queries [66, 109]. Further advanced techniques to query modification has been also considered by these approaches, such as query disambiguation [67] and relaxation [33, 34], which augment the query with further semantics. Traditional IR approaches [47] adopt query modification techniques yet adding/removing/modifying syntactic terms following classical Relevance Feedback. It is to note that, the same models and techniques used for query interpretation are usually used for the document representation in the annotation phase.

- *Search:* Traditional IR approaches are based on either the exact match or the best match theory [8]. SIR approaches focus on the best match model to search for resources that are exactly or partially relevant to the user's queries. They also differ from traditional approaches in the way they search for these resources. As they are based on semantic graphs, new search techniques have been introduced such as the graph traversal [82, 81, 93] and the sub-graph matching [36, 108].

- *Ranking:* Despite Boolean Retrieval models which do not consider scores for matched resources w.r.t. user's queries, traditional IR approaches mainly compute scores for terms and vectors of terms (e.g., using tf-idf, BM25, etc.). All these models are still used by SIR approaches, however the latter often apply them on concepts, nodes and graphs. SIR approaches also rely on semantic measures [80, 106] to compute these scores. It is to note that Search and Ranking phases are often merged in one single phase.

- *Presentation:* IR approaches based on the Boolean Retrieval model do not provide a ranking for the set of retrieved resources. Other traditional IR approaches [47, 85, 100] and SIR approaches [23, 32, 56, 81, 95] output, at the final stage of query processing, a ranked list of relevant resources.

TABLE 3.3 – Evaluation of the existing IR and SIR models (considering different approaches) w.r.t. the identified challenges.

		Challenge 2.1	*Challenge 2.2*	*Challenge 2.3*
		Relevant Granularity Levels	**Relevant Inter/Intra-Document Dependencies**	**Contextualized Query Answers**
Traditional IR Models		Partial	✗	✗
SIR Models	*Fully-fledged Approaches*	Partial	Partial	Partial
	SPARQL-based Approaches	Partial	Partial	Partial
	Graph-based Approaches	Partial	Partial	Partial
	Molecule-based Approaches	Partial	Partial	Partial

We evaluated the existing IR models based on the challenges previously mentioned in Sect. 3.1. The results are depicted in Table 3.3 where we used the same symbols as those for Sect. 2.2.3. In general, none of the existing IR models is able to answer the three challenges combined. In the following, we summarize their limitations w.r.t. each challenge:

- *Relevant Granularity Levels (Challenge 2.1):* Although current IR models (including traditional and SIR) search for relevant information contained in different granularity levels of the document structure, they do not provide in their results detailed information regarding the relevant parts of the documents. Instead, they provide the user with the whole documents that might contain many other irrelevant parts.

- *Relevant Inter/Intra-Document Dependencies (Challenge 2.2):* Traditional IR models do not consider the various dependencies between documents or parts of

the documents. This is because their underlying models focus on the presentation of documents as bag of words neglecting the relations between them. SIR models only focus on the lexical dependencies or domain-oriented links between the documents, thus partially considering these dependencies in their results.

- *Contextualized Query Answers (Challenge 2.3):* Existing IR models do not provide both structural and domain-specific context for query answers. Although partial contextual information could be extracted from the content of the provided resources (e.g., snippet of the relevant content) or from knowledge-graph relations between them (e.g., resources involving related semantic concepts), they still neglect structural information (e.g., document cross references, granularity levels). Thus, their retrieval results do not help the user in easily locating straightforward information and interpretation the results, especially in tracking cross document dependencies.

To sum up, existing SIR works [32, 34, 56, 81, 93, 95, 109] are suitable for several applications, however there are still some limitations w.r.t. the challenges we aim to cope with. This does also concerns the data models behind which completely neglect the structural information of the documents (e.g., parts of the documents) and the various dependencies between these structural parts (e.g., references). We discussed in Chapter 2, the advantages of adopting a tightly coupled semantic graph to represent the collective knowledge of a heterogeneous document corpus within a SIR system supporting the structural and the domain-specific dimensions, and the coupling between them. Thus, we consider adopting such a model as a prerequisite to deal with the identified challenges. A naive solution would be to apply one of the existing approaches to SIR on our tightly coupled semantic graph considering the four stages of the classical pipeline for query processing. In this thesis, we will not focus on query interpretation since we consider that relying on existing techniques would be sufficient to tackle the main problem. For instance, natural language queries, NLP and semantic annotations techniques [26] are convenient in our context for non computer expert users (such as actors of the AEC industry). Although visual SPARQL-based approaches [10, 45, 46, 90] and query re-writing or refinement techniques [3, 33, 34] aim at improving queries for users with limited technological skills, however they still require a minimum level of knowledge on the background data model's structure or the SPARQL language respectively. The main focus of this Chapter is to consider the identified challenges while searching, ranking and presenting relevant parts of our tightly coupled semantic graph w.r.t. a given user's query. Existing graph-based SIR approaches [81, 108, 93] do not provide an exhaustive solution since their search and ranking strategy do not consider the different dimensions of our graph. Most importantly, the presentation of their results relies on node-based query answers representing the whole documents or web pages. Molecule-based SIR approaches [27, 28, 30, 36, 69] take the

advantage of relying on well-defined sub-graphs, called RDF molecules, for different purposes (e.g., tracking RDF provenance, stream reasoning, federated SPARQL queries). Along the same lines, we consider that (i) defining a novel molecule data structure that copes with the characteristics of our tightly coupled semantic graph, and (ii) adapting current search, ranking and presentation strategies based on the proposed data structure would respond to the three identified challenges. In other words, this would generate meaningful augmented query answers that further consider the documents' structure and their inter and intra-document dependencies, and provide users with helpful information.

3.3 Hybrid Molecules for Tightly Coupled Semantic Graphs

In this section, we first give a brief overview on the motivations behind introducing a novel data structure, called *hybrid molecule*, on the basis of a tightly coupled semantic graph that we introduced in Chapter 2 (See Sect. 2.3.3). We then provide a formal definition for this structure. We also use Fig. 3.4, which depicts our motivating scenario (See Fig. 3.1) in a tightly coupled semantic graph, to provide illustrative examples.

3.3.1 Overview

As we explained in Chapter 2, given a heterogeneous document corpus, one important feature of its tightly coupled semantic graph stands in the coupling information between hybrid components i.e., the hybrid edges between components of the structural-based dimension and those of the domain-specific dimension. We consider that defining a novel data structure that emphasizes the hybrid information in a meaningful (well-defined) sub-graph is crucial for the use of this graph in several applications (such as clustering, query processing, etc.) and for the exploitation of interesting information.

In this study, we will focus on the use of this sub-graph structure for query processing. For instance, Fig. 3.5 shows an example of an extract of the tightly coupled semantic graph presented in Fig. 3.4 with a focus on the hybrid information involved between a domain-specific instance of *SolarFactor* (i.e., *solarfactor1* which label is *SW*) and a structural-based instance containing the value ("$SW = 0.45$") of solar factor (i.e., a table cell in the section entitled *"2) Characteristics of Window Frames"* of d_2), together with their domain-specific context (i.e., information related to *SolarFactor* concept), and their structural-based one (i.e., granularity levels of the documents, references, etc.). Components of this sub-graph throws back to elements of the desired contextualized query answer depicted in Fig 3.2 (See Sect. 3.1).

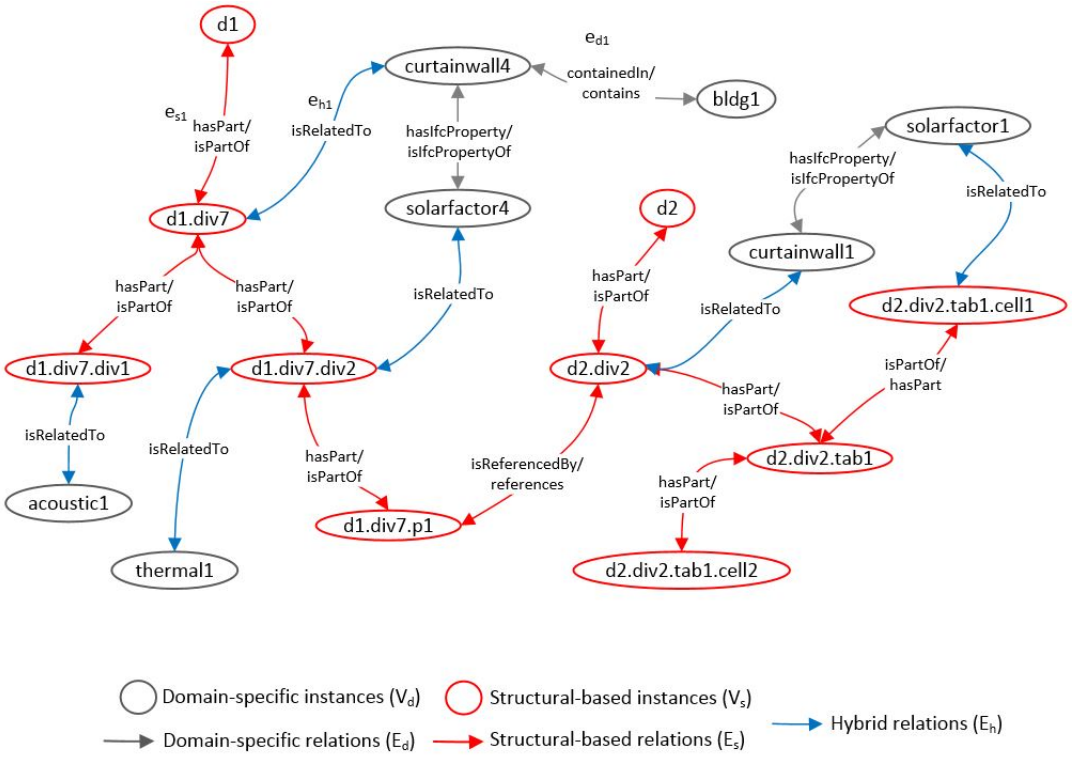


FIGURE 3.4 – Extract of a tightly coupled semantic graph $\mathcal{G}_{\delta_1}$ representing the two heterogeneous documents in Fig. 3.1.

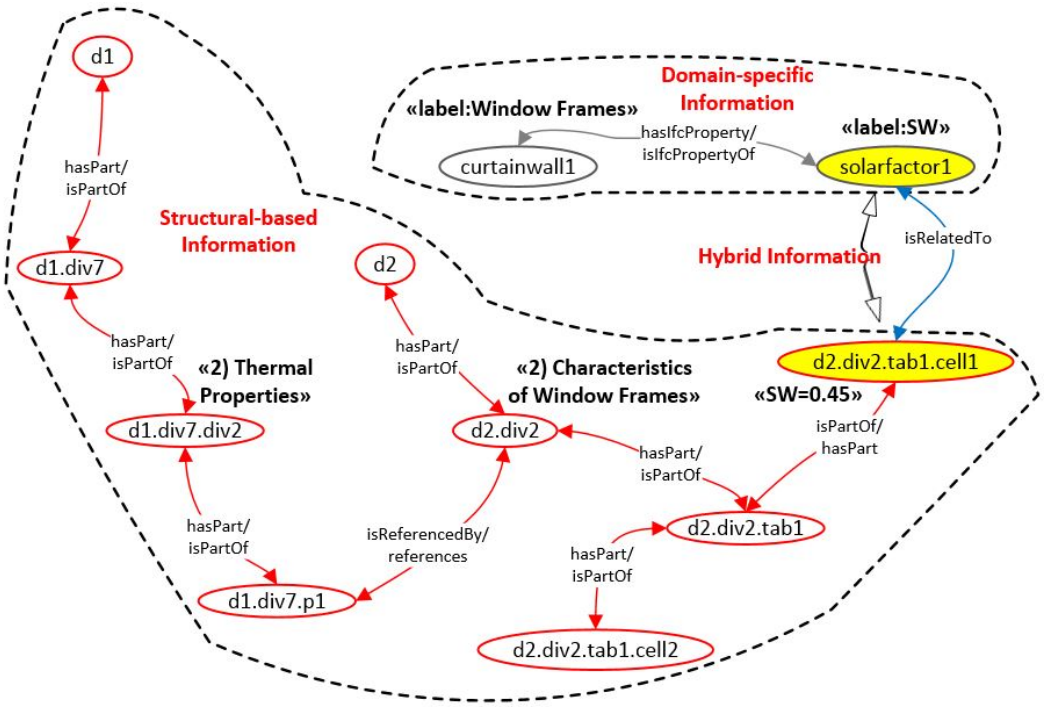


FIGURE 3.5 – Extract of a sub-graph of G_{δ_1} involving hybrid information with its contextual domain-specific information and structural-based one.

3.3.2 Hybrid Molecules

In this section, we introduce hybrid molecules which we build upon the definitions of molecules in the literature [27, 28, 30, 36, 69], yet regardless of the serialization technology. Molecules are sub-graphs of connected nodes. They are extracted from an initial graph using a decomposition function. We propose adjusting the decomposition to better cope with our tightly coupled semantic graph (See Sect. 2.3.3) and provide meaningful sub-graphs i.e., sub-graphs encapsulating a core information and a meaningful context that covers both structural and domain-specific features of a given heterogeneous document corpus. We formally define a hybrid molecule as follows:

Definition 4 (*Hybrid Molecule*)**.** Given the instances graph $\mathcal{G}_\delta$ describing a heterogeneous document corpus δ, we define a hybrid molecule $m(e_h, V_m, E_m, \Delta_{m_{max}}, w_m, f_{W_m})$, also denoted by $m \in M$, as a sub-graph decomposition result from the initial graph $\mathcal{G}_\delta$ based on a coupling between a domain-specific node and its related structural-based node, where:

- $M = d_{E_h}(\mathcal{G}_\delta)$ is the set of hybrid molecules representing sub-graphs obtained from the decomposition function d_{E_h}. This function splits the initial graph $\mathcal{G}_\delta$ into molecules whenever a hybrid edge $e_h \in E_h$ is identified, such that each molecule $m \in M$ has a unique $e_h \in E_h$.

  For instance, $d_{E_h}(\mathcal{G}_{\delta_1}) = \{m_1, m_2, m_3, m_4, m_5, m_6\}$ decomposes the graph $\mathcal{G}_{\delta_1}$, depicted in Fig. 3.4, into six different molecules since there are six different hybrid edges identified in $\mathcal{G}_{\delta_1}$ (See Fig. 3.6 $\rightarrow$ Fig. 3.11).

- $e_h \in E_h$ is the hybrid edge identifying the molecule m. Also, we refer to e_h as the *core* of the molecule m. We denote by $e_h.v_d \in V_d$ the domain-specific node of the molecule's core and $e_h.v_s \in V_s$ the structural-based node of the molecule's core.

  For instance, $e_{h_1} = (curtainwall4, d1.div7)$ is the core of the molecule m_1 (See. Fig. 3.6) where $e_{h_1}.v_d$ refers to the domain-specific node *curtainwall4* and $e_{h_1}.v_s$ refers to the structural-based node *d1.div7*.

- $V_m \subseteq (V_s \cup V_d)$ is the subset made of domain-specific and structural-based nodes forming m.

  For instance, *solarfactor4* and *bldg1* are examples of domain-specific nodes $v_d \in V_{m_1}$. *d1.div7.div1* and *d1.div7.p1* are examples of structural-based nodes $v_s \in V_{m_1}$ (See Fig. 3.6).

- $E_m \subseteq (\{e_h\} \cup E_d \cup E_s)$ is the subset made of the core edge, domain-specific and structural-based edges forming m and linking nodes $v \in V_m$.

  For instance, $e_{h_1} = (curtainwall4, d1.div7)$, $e_{d_1} = (curtainwall4, bldg1)$, and $e_{s_1} = (d1, d1.div7)$ are examples of hybrid, domain-specific, and structural-based edges respectively where $e_{h_1}, e_{d_1}, e_{s_1} \in E_{m_1}$ (See Fig. 3.6).

- $\Delta_{m_{max}}$ is the maximum distance in a molecule m that corresponds to the largest of a set of shortest paths from a node $v \in V_m$ to its closest molecule's core vertex (i.e., $e_h.v_d$ or $e_h.v_s$), such that:

$$\Delta_{m_{max}} = \max_{v \in V_m}(\min(shortestPath(v, e_h.v_s), shortestPath(v, e_h.v_d))), \quad (3.1)$$

  Where, *shortestPath* refers to the minimum number of edges between two nodes regardless of the weights of these edges.

  For instance, $\Delta_{m_{1max}} = 5$, which comes down to *shortestPath(d1.div7, d2.div2.tab1.cell1)* and *shortestPath(d1.div7, d2.div2.tab1.cell2)* given that both *d2.div2.tab1.cell1* and *d2.div2.tab1.cell2* are the farthest nodes to the core e_{h_1}, more precisely to its structural-based node $e_{h_1}.v_s$ i.e., *d1.div7* (See Fig. 3.6).

- w_m is the overall weight of the molecule m such that $w_m \in W_m$, where W_m is the set of molecules' weights.

- $f_{W_m} : M \rightarrow W_m$ is the molecule weight mapping that assigns the weight $w_m \in W_m$ to the molecule $m \in M$.

  The molecule weight mapping f_{W_m} is used in the ranking process of query answers. It is detailed in Sect. 3.4.4.3.

To sum up, the core of a hybrid molecule holds the molecule's central information as it is where the domain specific knowledge is anchored to a structural component in the document corpus. The rest of the molecule's nodes and edges augments the core with additional relevant information. For instance, in Fig 3.6, $e_{h_1} = (curtainwall4, d1.div7)$ is the core of the molecule m_1. *d1.div7* contains relevant information on the curtain walls (exterior facades). Other structural-based components (e.g., $e_{s_1} = (d1, sect1)$) and domain-specific ones (e.g., $e_{d_1} = (curtainwall4, bldg1)$) in V_{m_1} and E_{m_1} provide m_1 with further useful information (i.e., *d1.div7* is part of document d_1 and *bldg1* contains *curtainwall4*). Table 3.4 gives further examples for the 6 molecules.

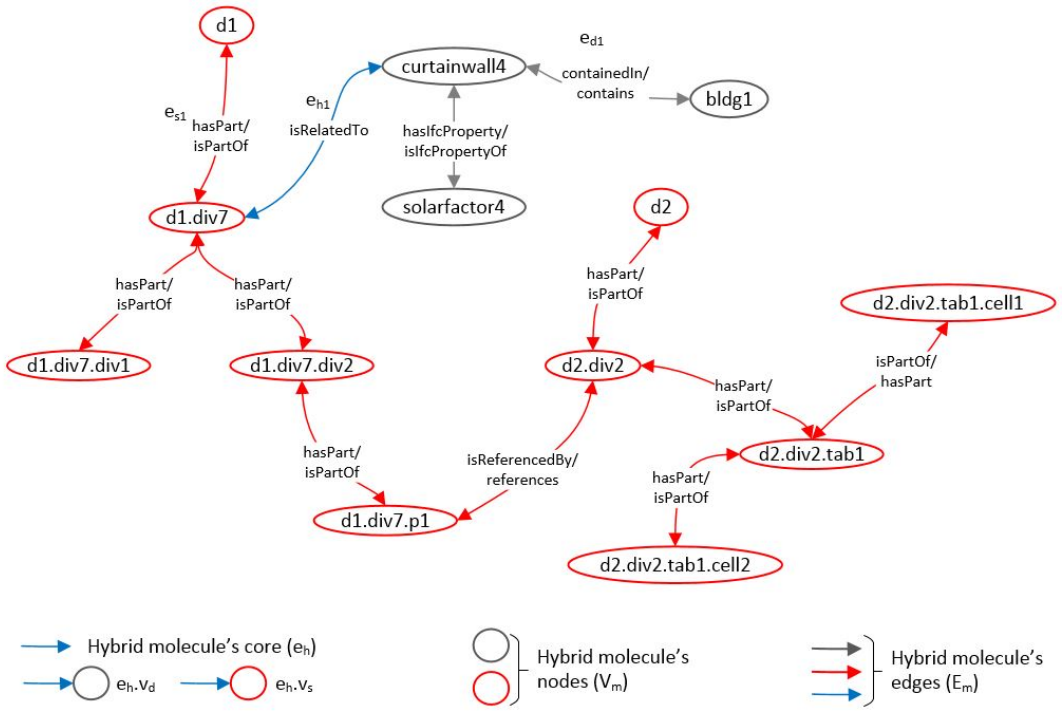


FIGURE 3.6 – Example of a Hybrid molecule m_1 extracted from $\mathcal{G}_{\delta_1}$ depicted in Fig. 3.4.

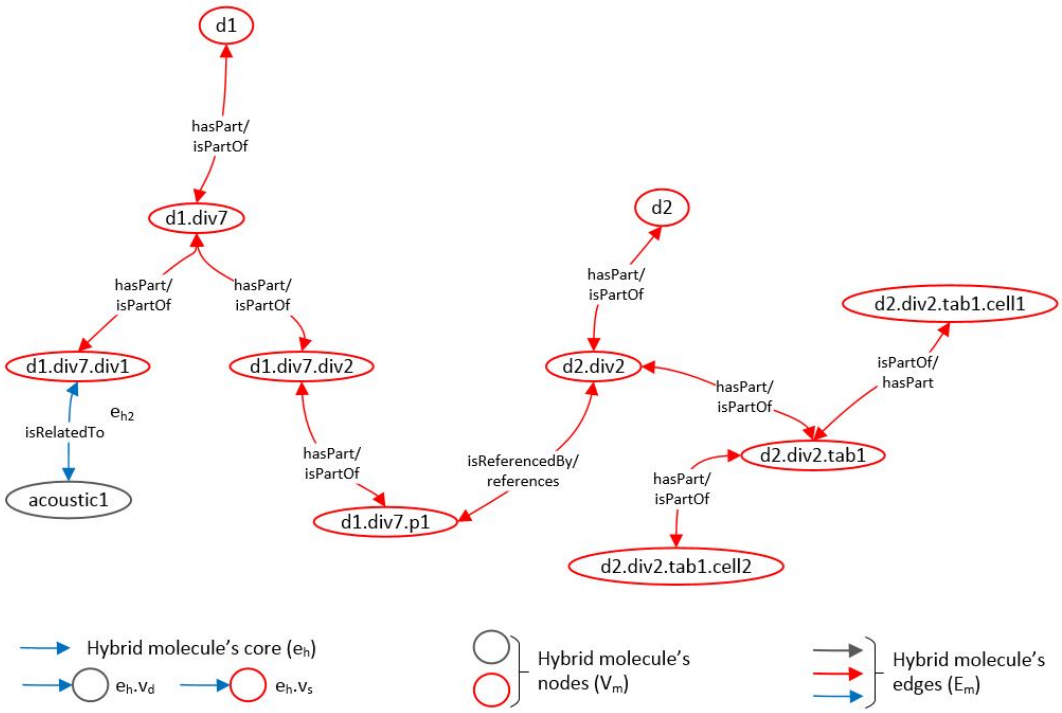


FIGURE 3.7 – Example of a Hybrid molecule m_2 extracted from $\mathcal{G}_{\delta_1}$ depicted in Fig. 3.4.

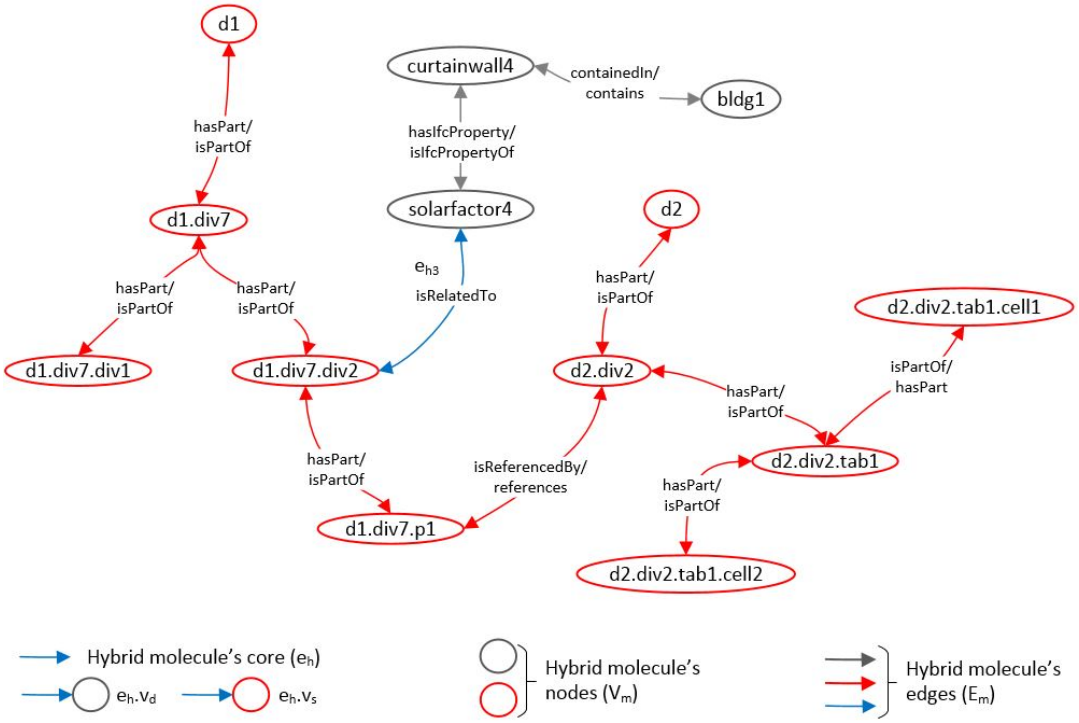


FIGURE 3.8 – Example of a Hybrid molecule m_3 extracted from $\mathcal{G}_{\delta_1}$ depicted in Fig. 3.4.

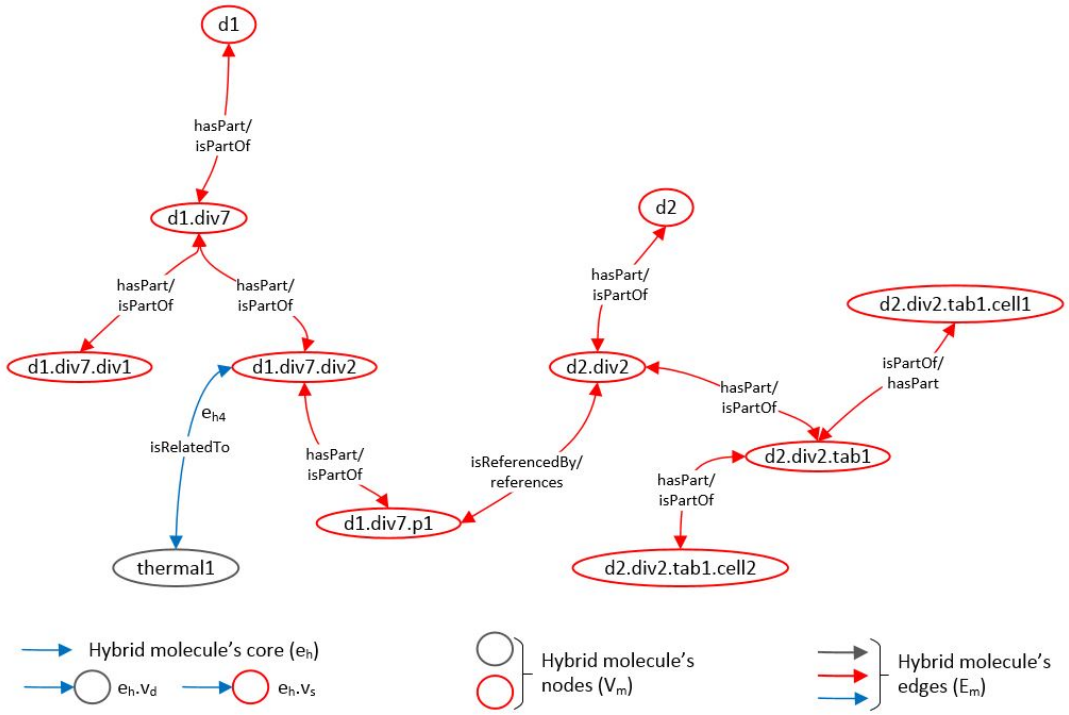


FIGURE 3.9 – Example of a Hybrid molecule m_4 extracted from $\mathcal{G}_{\delta_1}$ depicted in Fig. 3.4.

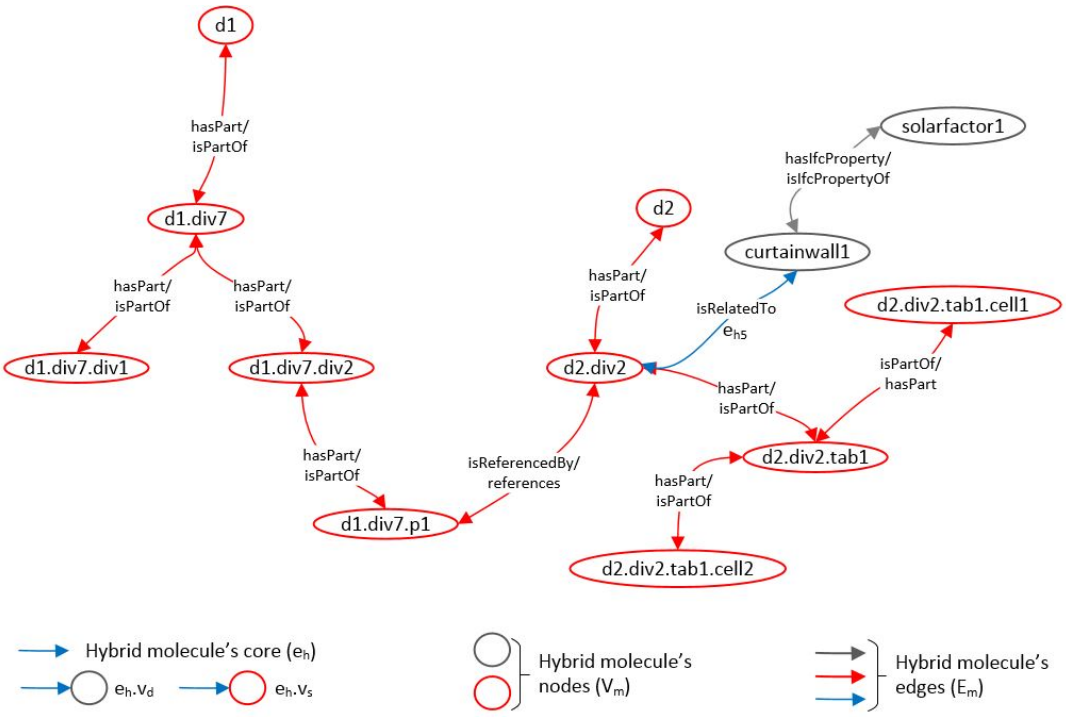


FIGURE 3.10 – Example of a Hybrid molecule m_5 extracted from $\mathcal{G}_{\delta_1}$ depicted in Fig. 3.4.

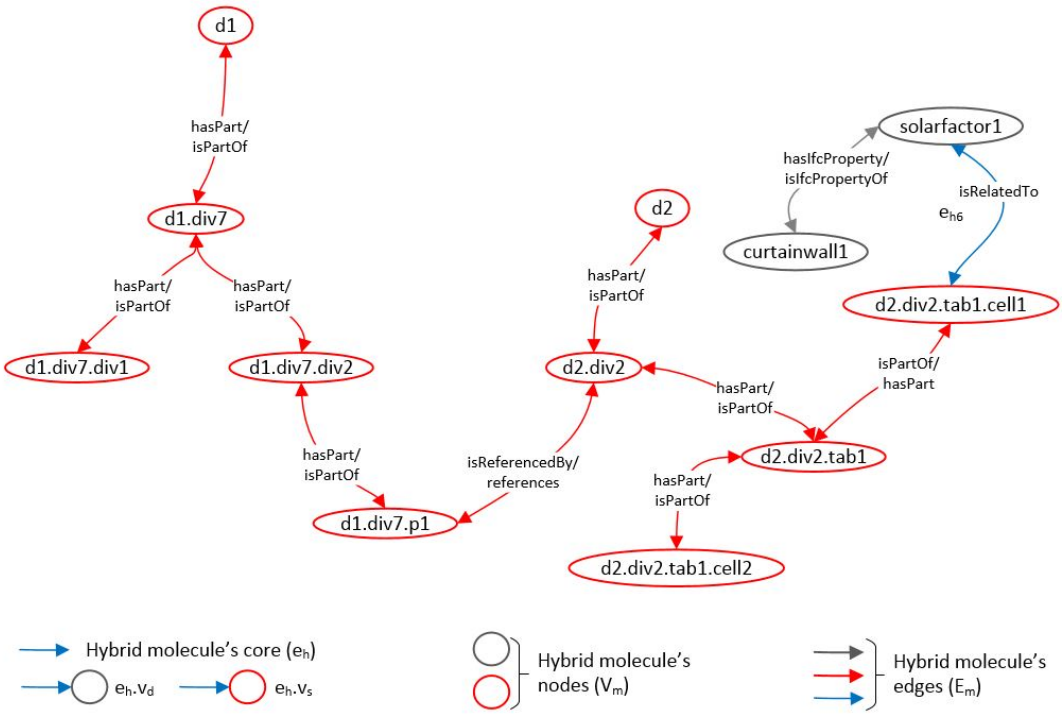


FIGURE 3.11 – Example of a Hybrid molecule m_6 extracted from $\mathcal{G}_{\delta_1}$ depicted in Fig. 3.4.

TABLE 3.4 – Core information of molecules extracted from $\mathcal{G}_{\delta_1}$ of Fig 3.4 together with examples of their structural-based and domain-specific contexts.

Hybrid Molecules	Core Information	E.g. of Additional Structural-based Information	E.g. of Additional Domain-specific Information
m_1	e_{h_1} Information on the curtain wall *curtainwall4* in *d1.div7*	*d1.div7* is part of document d_1	*curtainwall4* is contained in *bldg1*
m_2	e_{h_2} Information on acoustic properties *acoustic1* in *d1.div7.div1*	*d1.div7.div1* is part of document d_1	(not depicted in Fig. 3.7)
m_3	e_{h_3} Information on the solar factor *solarfactor4* in *d1.div7.div2*	*d1.div7.div2* is part of document d_1	*solarfactor4* is property of *curtainwall4*
m_4	e_{h_4} Information on thermal properties *thermal1* in *d1.div7.div2*	*d1.div7.div2* is part of document d_1	(not depicted in Fig. 3.9)
m_5	e_{h_5} Information on the curtain wall *curtainwall1* in *d2.div2*	*d2.div2* is part of document d_2	*curtainwall1* has property *solarfactor1*
m_6	e_{h_6} Information on the solar factor *solarfactor1* in *d2.div2.tab1.cell1*	*d2.div2.tab1.cell1* is part of document d_2	*solarfactor1* is property of *curtainwall1*

3.4 Query Processing over a Heterogeneous Document Corpus

This section provides an overview on a query processing pipeline over a heterogeneous document corpus. Although we provide a general algorithm including a bunch of other algorithms for each module of the whole pipeline, we particularly focus on the search and ranking modules.

3.4.1 Overview

Our aim is to retrieve relevant information from a given heterogeneous document corpus w.r.t. to non computer expert users' queries. To do this, we propose a query processing strategy that relies on:

- A tightly coupled semantic graph representing the heterogeneous document corpus, so as to take advantage from the richness of this model and the collective knowledge embedded in it,
- Hybrid molecules defined in Sect. 3.3, which provide a novel data structure for query answers made of core information (i.e., hybrid couples) and additional contextual information from both structural and domain-specific dimensions,
- Adapted algorithms that are able to identify components of the proposed hybrid molecules and construct them progressively, starting from the query interpretation stage until the graph-based search, ranking and presentation strategies handling the characteristics of these molecules.

Our proposal aligns with the previously mentioned challenges (See Sect. 3.1) since the hybrid molecules query answers provide users with (i) structural granularity parts of the documents associated with relevant domain-specific information (from the core), (ii) relevant information on the various dependencies between the documents and parts of the document (from structural and domain-specific contextual information associated to the core), and (iii) a helpful frame to interpret the results based on a meaningful sub-graph structure.

Fig. 3.12 illustrates our proposal and recalls the classical pipeline for query processing in an IR system (See Fig. 3.3), yet integrating the notion of Hybrid Molecules (HM) in all the stages and relying on a tightly coupled semantic graph for representation of the heterogeneous document corpus, and on *LinkedMDR* and a domain-specific ontology (for its Pluggable Layer) as the underlying knowledge structure. This corresponds to the upper layer of *FEED2SEARCH* (See Sect. 1.4.1). The input of our proposed strategy to query processing is a natural language query (expressed as plain text). The final output comes down to relevant hybrid molecules query answers representing parts of documents or documents with additional contextual information. Since the end users are non computer experts, these molecules are presented in a SERP-like results.

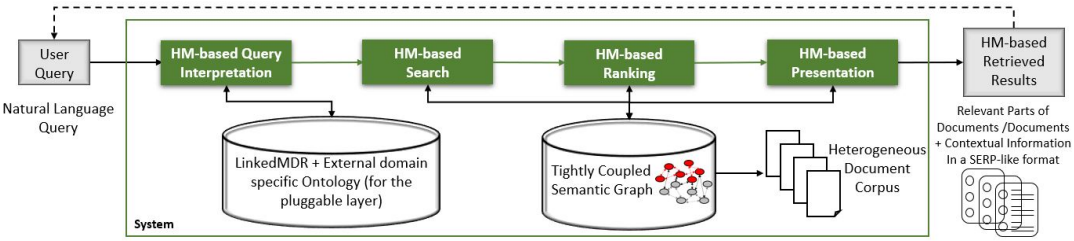


FIGURE 3.12 – A Hybrid Molecule (HM)-based pipeline for query processing in IR.

3.4.2 Overall Algorithm

In this section, we present the overall algorithm of our proposed HM-based query processing (See Algorithm 2) over a heterogeneous document corpus.

Algorithme 2 : HM Query Processing

Inputs : User's natural language query q; External domain-specific ontology $\mathcal{O}_{\mathcal{D}}$; Underlying background ontology *LinkedMDR*; Tightly coupled semantic graph $\mathcal{G}_{\delta}$ describing a given heterogeneous document corpus δ; Set of search parameters *search_params*; Set of ranking parameters *rank_params*.

Output : List of hybrid molecule-based answers M_{out} printed in a SERP-like format.

```
// **STEP 1** HM-based Query Interpretation
1 I_d_out ← HM_QueryInterpretation(q, O_D, LinkedMDR) ;   // extracts LinkedMDR domain-specific
   instances I_d_out from the query using O_D and LinkedMDR's domain-specific layer
// **STEP 2** HM-based Search
2 M_out ← HM_Search(I_d_out, G_δ, search_params) ;   // applies a graph-based search strategy on G_δ
   using search_params and generates a set of possibly relevant Hybrid Molecules w.r.t.
   the user's query q
// **STEP 3** HM-based Ranking
3 M_out ← HM_Ranking(M_out, rank_params) ; // applies a ranking strategy on the Hybrid Molecule
   results using rank_params and generates a ranked list of Hybrid Molecules
// **STEP 4** HM-based Presentation
4 HM_Presentation(M_out) ;   // transforms and prints the Hybrid Molecule-based answers in a
   SERP-like format
5 return ;
```

The algorithm takes as input: (i) a natural language query q expressed by a non computer expert user (such as an actor in the AEC industry) in plain text, (ii) the external domain-specific ontology $\mathcal{O}_{\mathcal{D}}$ (See Sect 2.3.3.1) that was plugged into *LinkedMDR*'s domain-specific layer (See Sect. 2.3.2.3), (iii) *LinkedMDR* ontology (See Sect. 2.3.2), (iv) a tightly coupled semantic graph $\mathcal{G}_{\delta}$ generated based on a heterogeneous document corpus δ, an external domain-specific ontology $\mathcal{O}_{\mathcal{D}}$ and the infrastructure of *LinkedMDR* (See Sect. 2.3.3), (v) a set of search parameters *search_params* used to configure the graph-based search strategy, and (vi) a set of ranking parameters *rank_params* used to configure the ranking strategy of the query answers. The final output is a list of ranked molecule-based query answers presented in a SERP-like format. This is done following four major steps. They are summarized as follows:

- Step 1 (line 1): It calls *HM_QueryInterpretation* algorithm (See Algorithm 3) described in Sect. 3.4.2.1. The main purpose is to extract domain specific instances I_{d_out} from the user's natural language query q. This is done by using the semantics of $\mathcal{O}_{\mathcal{D}}$ and the infrastructure provided by *LinkedMDR*'s domain-specific layer.

- Step 2 (line 2): It calls *HM_Search* algorithm (See Algorithm 4) described in Sect. 3.4.2.2. The domain-specific instances I_{d_out}, which were extracted from the previous step, are now used as input for this algorithm to start a graph-based search strategy on the graph $\mathcal{G}_{\delta}$. The main purpose to search for possibly relevant hybrid molecules M_{out} based on previous knowledge (i.e., I_{d_out}) and *search_params*. The latter provide the configurations required for the graph traversal.

- Step 3 (line 3): It calls *HM_Ranking* algorithm (See Algorithm 5) described in Sect. 3.4.2.3. The hybrid molecules M_{out}, which were extracted from the previous step, are now used as input for this algorithm. The main purpose is to rank the hybrid molecules M_{out} and provide a list of ranked hybrid molecule-based

query answers after applying a ranking strategy using *rank_params*. The latter provide the configurations required for assigning weights or scores to the hybrid molecules.

- Step 4 (line 4): It calls *HM_Presentation* algorithm (See Algorithm 6) described in Sect. 3.4.2.4. The hybrid molecules M_{out}, which were ranked from the previous step, are now used as input for this algorithm. The main purpose is to transform and present these molecules in SERP-like results for them to be more understandable by the user.

The notion of hybrid molecules intervenes in the four steps. Starting from the query interpretation, domain-specific elements are identified from the query, which will then match domain-specific nodes in the graph. More specifically, the matching parts of the graph come down to domain-specific node parts of the cores of possibly relevant hybrid molecules. The search, ranking and presentation phases rely on our proposed definition of hybrid molecules to construct them, rank them and present them conveniently.

3.4.2.1 Hybrid Molecule-based Query Interpretation

This section provides details on the hybrid molecule-based query interpretation phase (See Algorithm 3). More specifically, it describes *HM_QueryInterpretation* function used by Algorithm 2.

HM_QueryInterpretation converts the user's natural language query q into an understandable format that is compatible with the system's internal knowledge structure (i.e., the tightly coupled semantic graph). Since the user's background corresponds to knowledge related to a domain-specific application, we consider that elements describing the domain-specific dimension of the graph are the best to interpret the user's needs. Thus, *HM_QueryInterpretation* extracts, from q, domain-specific instances I_{d_out} that could later match the graph $\mathcal{G}_\delta$, specifically domain-specific nodes of its hybrid edges. This is done using the same techniques offered at the middleware indexing layer and used during the generation of the domain-specific parts of the graph (See Sect. 2.3.3.3):

- Step 1 (line 1): It generates semantic annotations ($Output_d$) describing domain-specific information identified from the textual content of the query q using the semantics of the external domain-specific ontology $\mathcal{O}_\mathcal{D}$.
- Step 2 (line 2): It converts the generated descriptors of the previous step into domain-specific instances (I_{d_out}) of *LinkedMDR* using the semantics of *LinkedMDR*'s domain-specific layer.

3.4.2.2 Hybrid Molecule-based Search

This section provides details on the hybrid molecule-based search phase (See Algorithm 4). More specifically, it describes the *HM_Search* function used by Algorithm 2.

Algorithme 3 : HM_QueryInterpretation

Inputs : User's natural language query q; External domain-specific ontology $\mathcal{O}_\mathcal{D}$; Underlying background ontology *LinkedMDR*.
Output : List of *LinkedMDR* domain-specific instances I_{d_out}.

```
// **STEP 1** using techniques for automatic semantic annotations
1 Output_d ← middleware(q, O_D) ;   // generates semantic annotations as output from the textual
   content of the natural language query q
// **STEP 2** using our tailored converters
2 I_d_out ← middleware(Output_d, LinkedMDR) ;          // converts the query output into LinkedMDR
   domain-specific instances I_d_out relying on the semantics of LinkedMDR's domain-specific
   layer
3 return I_d_out;
```

HM_Search applies a graph-based search strategy on $\mathcal{G}_\delta$ in order to retrieve relevant hybrid molecules. This is done following two major steps:

- Step 1 (lines 1-5): It matches domain-specific nodes in the graph $\mathcal{G}_\delta$ based on previously extracted domain-specific instances (I_{d_in}) in the query interpretation phase. It uses concept matching i.e., for each instance $i \in I_{d\_in}$, it searches for all domain-specific instances V_{d_temp} in $\mathcal{G}_\delta$ where each $v_{d\_temp} \in V_{d\_temp}$ the same concept type as i (line 3). These instances are then appended to V_{d_in} (line 4) and considered as start nodes for the graph traversal of the next step.
- Step 2 (line 6): It traverses the graph $\mathcal{G}_\delta$ starting from nodes in V_{d_in} and begins to search for relevant hybrid molecules using *search_params*. The underlying algorithm, which we call *HM_CSA*, is inspired by the CSA graph-based strategy [24], and constructs hybrid molecules progressively while traversing the graph $\mathcal{G}_\delta$. It is detailed in Sect. 3.4.3.

Algorithme 4 : HM_Search

Inputs : *LinkedMDR*'s domain specific instances I_{d_in}; Tightly coupled semantic graph $\mathcal{G}_\delta$ describing a given heterogeneous document corpus δ; Set of search parameters *search_params*.
Output : List of hybrid molecules M_{out}.

```
// **STEP 1** identifying domain-specific start nodes in G_δ
1 V_d_in ← ∅ ;          // Initializes the set of domain-specific start nodes used to trigger a
   graph-based search on G_δ
2 foreach i ∈ I_d_in do
3     V_d_temp ← matchQuery(i, G_δ) ;       // identifies a set of domain-specific instances of G_δ
       matching instance i
4     V_d_in ← V_d_in ∪ V_d_temp ;   // appends domain-specific identified nodes to the set of start
       nodes V_d_in
5 end
// **STEP 2** searching for possibly relevant hybrid molecules in G_δ
6 M_out ← HM_CSA(V_d_in, G_δ, search_params) ;    // triggers a graph-based traversal algorithm on
   the graph G_δ starting from nodes in V_d_in and using search_params to generate a set of
   possibly relevant Hybrid Molecules
7 return M_out;
```

3.4.2.3 Hybrid Molecule-based Ranking

This section provides details on the hybrid molecule-based ranking phase (See Algorithm 5). More specifically, it describes the *HM_Ranking* function used by Algorithm 2.

HM_Ranking assigns scores or weights to the retrieved hybrid molecules from the search stage. It ranks them according to the computed weights. This is done using two major steps:

- Step 1 (lines 1-3): It assigns a score or weight w_m to each hybrid molecule $m \in M_{in}$ retrieved from the previous stage using *rank_params*. The *HM_Score* function uses the weight mapping f_{W_m} associated to the hybrid molecule's definition (See Sect. 3.3.2) and is detailed in Sect. 3.4.4.
- Step 2 (line 4): It ranks the weighted hybrid molecules in descending order following their individual weights and generates a list of ranked hybrid molecules M_{out}.

Algorithme 5 : HM_Ranking

Inputs : Set of Hybrid Molecules M_{in}; Set of ranking parameters *rank_params*.
Output : List of ranked hybrid molecules M_{out}.

```
// **STEP 1** scoring the hybrid molecules
1 foreach m ∈ M_in do
2 |   m ← HM_Score(m, rank_params) ;        // assigns a score (weight) w_m to each molecule
  |     m ∈ M_in using rank_params
3 end
// **STEP 2** ranking the hybrid molecules
4 M_out ← rankMolecules(M_out) ;     // rank the hybrid molecules in descending order following
   their individual scores
5 return M_out;
```

3.4.2.4 Hybrid Molecule-based Presentation

This section provides details on the hybrid molecule-based presentation phase (See Algorithm 6). More specifically, it describes the *HM_Presentation* procedure used by Algorithm 2.

HM_Presentation aims at presenting the ranked list of hybrid molecules generated by the previous ranking stage in SERP-like results. This is to translate the hybrid molecule query answers into a more understandable output for non computer expert users (such as the output presented in Fig. 3.2, in Sect. 3.1). Regardless of the GUI behind, Algorithm 6 provides the following strategy in order to present a hybrid molecule $m \in M_{in}$:

- Step 1 (line 2): It displays core information i.e., values of the core's structural-based node $e_h.v_s$ and the core's domain-specific node $e_h.v_d$. Additional information on the structural-based node $e_h.v_s$ is also provided such as the corresponding document name and a link to it.
- Step 2 (line 3): It displays domain-specific contextual information i.e., the context of the core's domain-specific node $e_h.v_d$ made of connected domain-specific nodes and edges.

- Step 3 (line 4): It displays structural-based contextual information i.e., the context of the core's structural-based node $e_h.v_s$ made of connected structural-based nodes and edges.

Algorithme 6 : HM_Presentation

Input : Set of Ranked Hybrid Molecules M_{in};
Output : Printed hybrid molecules in SERP-like results.

```
1 foreach m ∈ M_in do
      // **STEP 1** printing core information
2     print(e_h) ;        // prints values of core's domain-specific and structural-based nodes
      // **STEP 2** printing domain-specific contextual information
3     print_ContextD(e_h.v_d) ;   // prints values and labels of connected domain-specific nodes
        and edges respectively in m
      // **STEP 3** printing structural-based contextual information
4     print_ContextS(e_h.v_s) ;   // prints values and labels of connected structural-based nodes
        and edges respectively in m
5 end
6 return ;
```

3.4.3 Hybrid Molecule-based Search by Constrained Spread Activation (CSA)

This section presents a novel graph-based search algorithm, *HM_CSA*, used during the Hybrid Molecule-based Search module (See Sect. 3.4.2.3). The main purpose of *HM_CSA* is to cope with the characteristics of a tightly coupled semantic graph $\mathcal{G}_\delta$ representing a heterogeneous document corpus δ, and retrieve relevant hybrid molecules from $\mathcal{G}_\delta$. We first provide some background on the CSA theory on which our proposed algorithm is build and the motivations behind our choice of this theory as a basis for our proposed algorithm, recall limitations of existing *CSA*-based algorithms, and present the main novelty of our proposal. We then describe *HM_CSA* in details with examples based on Fig. 3.4.

3.4.3.1 Constrained Spread Activation (*CSA*)

From the various graph-based search approaches (See Sect. 3.2.3.3), our proposed algorithm relies on a Breadth-First Search (BFS) graph traversal strategy[11], mainly the *CSA* algorithm [24]. *CSA* works by spreading out the activation from a set of start nodes to adjacent nodes progressively until predefined constraints are met. We consider *CSA* as a suitable search strategy for our proposed tightly coupled semantic graph as (i) it handles the heterogeneity of the graph since it can explore possibly relevant structural-based and domain-specific nodes located anywhere in the graph, (ii) it constructs progressively multiple target nodes from activated nodes, and (iii) it supports the incorporation of useful constraints, either at the beginning to select start nodes or at termination point to stop the spreading in the graph.

[11] BFS searches the graph data structure starting at chosen nodes in the graph of the neighbor nodes at the current depth prior to moving on to the nodes at the next depth level.

Although, in its standard form, *CSA* stands out as a simple yet effective solution in many IR applications [24, 25, 43, 81, 93], however, *CSA*-based algorithms still have some limitations w.r.t. the identified challenges (See Sect. 3.1):

- They output a ranked list of single nodes as query answers, which does not align with our main objective (i.e., to output hybrid-molecules as query answers),
- They weigh nodes and edges based on variant weight mapping strategies (e.g., cluster similarity measure[12] presented in [81]), however only adapted to domain-specific semantic graphs where nodes represent documents or web pages. This partially copes with the characteristics of our tightly coupled semantic graph as the latter involves heterogeneous nodes including structural-based ones belonging to different granularity levels of the documents and more sophisticated edges including hybrid edges, which may wrongly guide the activation through the graph towards undesired nodes.

Thus, we propose extending *CSA* to *HM_CSA* (See Algorithm 7) with the following advances w.r.t. existing *CSA*-based algorithms:

- It generates a list of hybrid molecules as query answers,
- It adapts the edge weights differently in order to handle the characteristics of desired hybrid molecules (i.e., hybrid core, structural-based nodes and edges, and domain-specific nodes and edges) and prepare them to be ranked conveniently for the ranking phase of the query processing.

3.4.3.2 *HM_CSA* Algorithm

The pseudo code of *HM_CSA* algorithm is presented in Algorithm 7. The input consists of (i) a set of domain-specific nodes V_{d_in} generated based on the output of the query interpretation module as they matched the user query (See Algorithm 4), (ii) a tightly coupled semantic graph $\mathcal{G}_\delta$ representing a heterogeneous document corpus δ, and (iii) a set of constraints parameters, *search_params*. The latter is made of pre-adjustment parameters (*search_params.preadjustment*) to be checked before each spread iteration occurs (e.g., a firing threshold F, and a maximum spread distance D from start nodes), post-adjustment parameters (*search_params.postadjustment*) to be checked by the end of each spread iteration (e.g., a maximum number of iterations I, and a maximum processing time T), and spread configurations (e.g., an activation percent decrease γ which imposes a decay on the propagation of the activation through the graph). The choice of these parameters is application-dependent. The output of the proposed algorithm consists of a list of relevant hybrid molecules M_{out}, where composing nodes are weighted based on their final activation values.

[12]Measures the similarity between two concepts based on the number of common relations with other concepts.

Algorithme 7 : HM_CSA

Inputs : Set of domain-specific start nodes V_{d_in}; Tightly coupled semantic graph $\mathcal{G}_\delta$; Constraint parameters *search_params*.
Output : List of relevant hybrid molecules M_{out}.

$V_{in} \leftarrow V_{d\_in}$; // set of activated nodes
$V_{out} \leftarrow \emptyset$; // set of spread nodes
$stopSpread \leftarrow false$; // boolean checking whether to stop CSA or not
while $(|V_{in}| > 0$ *AND* $!stopSpread)$ **do**
  // **STEP 1** Processing Firing Node
  $v_i \leftarrow getFiringNode(V_{in})$; // node with highest activation value
  $V_{in} \leftarrow V_{in} - \{v_i\}$; // removes firing node from V_{in}
  // **STEP 2** Checking Pre-adjustment Parameters
  if $(checkRestrictions(v_i, search\_params.preadjustment))$ **then**
    // **STEP 3** Exploring Neighbors
    $E_{ij} \leftarrow getNeighbors(v_i)$; // set of outgoing edges from v_i to the set of direct neighbors v_j
    // **STEP 4**Spreading out activation
    foreach $e_{ij} \in E_{ij}$ **do**
      // Using adapted edge weight mapping $f_{W_e}(e_{ij})$
      $\Delta_{input(v_j)} \leftarrow Output(v_i) \times f_{W_e}(e_{ij}) \times (1 - \gamma)$; // the contribution of neighbor v_i through e_{ij}. $Output(v_i) = Activation(v_i)$ i.e., the activation value of v_i
      $Input(v_j) \leftarrow Input(v_j) + \Delta_{input(v_j)}$; // input value of v_j
      $Output(v_j) \leftarrow normalize(Input(v_j))$; // output of v_j i.e., its activation value after normalization function
      // **STEP 5** Processing Neighbor
      if $(v_j \notin V_{in})$ **then**
        $V_{in} \leftarrow V_{in} \cup \{v_j\}$; // adds neighbor v_j to the set of activated nodes V_{in}
      else
        // Constructing and Processing Hybrid Molecules
        if $(v_j \in V_{out})$ **then**
          if $(isHybrid(e_{ij}))$ **then**
            $m_i \leftarrow createMolecule(e_{ij})$; // new molecule m_i
            $m_i \leftarrow appendFromMolecules(m_i, M_{out})$; // appends nodes and edges from existing molecules in M_{out} to the newly created molecule m_i
            $M_{out} \leftarrow M_{out} \cup \{m_i\}$; // add m_i to M_{out}
          else
            $M_{out} \leftarrow appendToMolecules(e_{ij}, M_{out})$; // appends current neighbor e_{ij} to existing molecules in M_{out}
          end
        end
      end
    end
    $V_{out} \leftarrow V_{out} \cup \{v_i\}$; // adds firing node v_i to V_{out} after activation's propagation is done
  end
  // **STEP 6** Checking Post-adjustment Parameters
  $stopSpread \leftarrow checkRestrictions(search\_params.postadjustment)$;
end
return M_{out};

In general, *HM_CSA* can be divided into two main parts which are detailed in the remainder of this section: (1) the implementation of the *CSA* theory (i.e., Algorithm 7 excluding red-highlighted section) with adapted weight mapping (i.e., the blue-highlighted section of Algorithm 7), and (2) the construction and processing of hybrid molecules (i.e., the red-highlighted section of Algorithm 7).

Implementation of the *CSA* Theory - We use two sets of nodes V_{in} and V_{out}. The former is fed with activated nodes as the spread activation processes through the graph while the second consists of spread nodes i.e., nodes which have activated

others and will go in the final output. At start time, V_{in} contains previously selected start nodes V_{d_in} where the activation value $Activation(v_i)$ of each node $v_i \in V_{in}$ is set to 1 (max value), and V_{out} is initially empty. The algorithm works by checking constraint parameters to decide whether to trigger or not a spread iteration where a firing node $v_i \in V_{in}$ propagates its activation to its neighbors. The same process repeats until V_{in} has no further nodes to process or post-adjustment constraints are met (line 4). This is done in six major steps described as follows:

- Step 1 (lines 5-6): It consists of removing a firing node v_i from V_{in} i.e., the node with the highest activation value. Note that, when many nodes have an equal highest activation value, *HM_CSA* selects the first node.

- Step 2 (line 7): It checks *search_params.preadjustment* i.e., pre-adjustment parameters to decide whether current firing node v_i is allowed to spread its activation or not to its neighbors. These constraints are met if v_i has a distance to start nodes less than or equal to the maximum spread distance D and an activation value higher than or equal to the firing threshold F.

- Step 3 (line 8): It explores direct neighbors of the firing node v_i through outgoing edges E_{ij} from v_i where each edge $e_{ij} \in E_{ij}$ connects v_i to its direct neighbor v_j.

- Step 4 (lines 9-13, 30): Firing node v_i spreads out its activation to each neighbor v_j. In other words, each explored node v_i contributes to the input of its neighbor v_j by $\Delta_{input(v_j)} \leftarrow Output(v_i) \times f_W(e_{ij}) \times (1-\gamma)$ (line 11), where $Output(v_i)$ is its current activation value (i.e., $Activation(v_i)$), $f_W(e_{ij})$ computes the weight $w_{e_{ij}}$ of the link e_{ij} connecting v_i to v_j (See Sect. 3.4.4.1) and $(1-\gamma)$ is the decay factor. Each $\Delta_{input(v_j)}$ is then added to the input of v_j i.e., $Input(v_j)$[13] (line 12). The actual activation value $Activation(v_j) = Output(v_j)$ of a node v_j is obtained by normalizing the sum of all the contributions it receives (line 13). *HM_CSA* uses a simple feature scaling function[14] to rescale values between 0 and 1. After having spread its activation, current node v_i is added to the set of spread nodes V_{out} (line 30) so it could be visited for molecules processing in future iterations.

- Step 5 (lines 14-15): It adds activated neighbor v_j to the set of activated nodes V_{in} if it is visited for the first time.

- Step 6 (line 32): It checks *search_params.postadjustment* i.e., post-adjustment parameters to decide whether to stop the spread iterations afterwards. These constraints are met (i.e., $stopSpread$ is $false$) if current spread iteration is less

[13] At start time, the input value $Input(v_j)$ is set to the initial activation value of v_j i.e., 1 for start nodes and 0 for others.

[14] $Output(v_j) = \frac{Input(v_j) - Input_{min}}{Input_{max} - Input_{min}}$ where $Input_{min}$, $Input_{max}$ are respectively the minimum and the maximum values of all nodes' inputs in the graph.

than the maximum number of allowed iterations I and the current processing time of the algorithm is less than the maximum processing time T.

Molecules Construction and Processing - One major added value of *HM_CSA* remains in the construction and processing of the hybrid molecules, as described in the red-highlighted section of Algorithm 7 (lines 17-26). We consider that all nodes that have spread out their activation i.e., contained in V_{out}, should take part in the final results in the form of hybrid molecules. One naive way to construct the hybrid molecules from the standard implementation of *CSA* and according to Definition 4 (See Sect. 3.3) is to post-process the nodes in V_{out}. However, this would require re-exploring connected nodes in V_{out} in order to construct the sub-graphs and would enormously increase the processing time of the algorithm. Instead, *HM_CSA* integrates the hybrid molecules construction process within the graph traversal while fetching and activating neighboring nodes.

Two cases arise when processing a neighbor $v_j \in V_{out}$ connected to v_i through an edge e_{ij}:

1. **Neighboring edge e_{ij} is hybrid** - i.e., $e_{ij} = e_h$ (lines 19-22): a new molecule m_i is constructed (according to Definition 4) with e_{ij} being the core of this molecule (line 20). Since a molecule should be also made of contextual information, *HM_CSA* appends connected nodes (except those connected through hybrid edges[15]) from existing molecules in M_{out}, created in previous iterations, whenever one of the core's nodes (i.e., $e_h.v_s$ or $e_h.v_d$) is found in these molecules (line 21). Algorithm 8 provides details on this operation. Given a newly created hybrid molecule m_{new} with e_h being its core and a set of previously created hybrid molecules M, the algorithm fetches each hybrid molecule $m_k \in M$ (line 1). Whenever the structural-based core's node $e_h.v_s$ is found in the hybrid molecule m_k, the algorithm uses σ_s, a structural-based selection operator, to only select the structural-based nodes and edges of m_k and adds these components to m_{new} (line 2-4). Likewise, whenever the domain-specific core's node $e_h.v_d$ is found in a hybrid molecule m_k, the algorithm uses σ_d, a domain-specific selection operator, to only select the domain-specific nodes and edges of m_k and adds these components to m_{new} (line 5-7). This results in an updated hybrid molecule m_{new} with related contextual information from possibly appended structural-based and domain-specific elements.

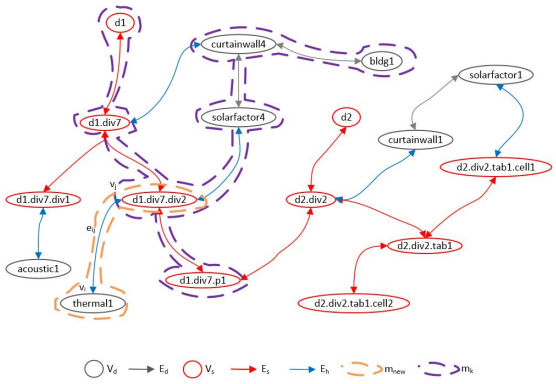
   For instance, Fig. 3.13a shows an example of a new molecule m_{new} with $e_{ij} = e_h$ = *(thermal1, d1.div7.div2)* being its core, and an existing molecule $m_k \in M$, where $e_h.vs$ = *d1.div7.div2* i.e., the structural-based core's node of m_{new} is found. Fig. 3.13b shows connected structural-based nodes, that are selected from m_k using the operator σ_s, which are then appended to m_{new} in Fig. 3.13c. The resulting hybrid molecules are depicted in Fig. 3.13d.

[15] The only hybrid edge of the molecule corresponds to its core.

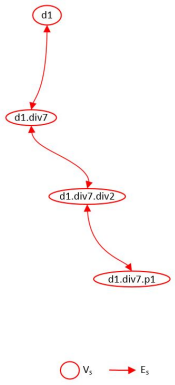
Algorithme 8 : appendFromMolecules

Inputs : Newly created hybrid molecule m_{new} with core e_h; Set of existing hybrid molecules M.
Output : Updated hybrid molecule m_{new} with possibly new elements appended to it.

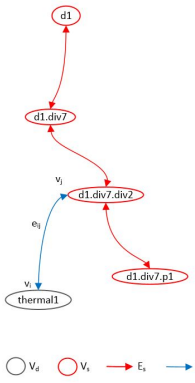
```
1  foreach m_k ∈ M do
2      if (e_h.v_s ∈ m_k) then
3          m_new ← m_new ∪ (σ_s(m_k)) ;  // appends only structural-based nodes and edges of m_k to
             m_new using σ_s as structural-based selection operator
4      end
5      if (e_h.v_d ∈ m_k) then
6          m_new ← m_new ∪ (σ_d(m_k)) ;  // appends only domain-specific nodes and edges of m_k to
             m_new using σ_d as a domain-specific selection operator
7      end
8  end
9  return m_new;
```



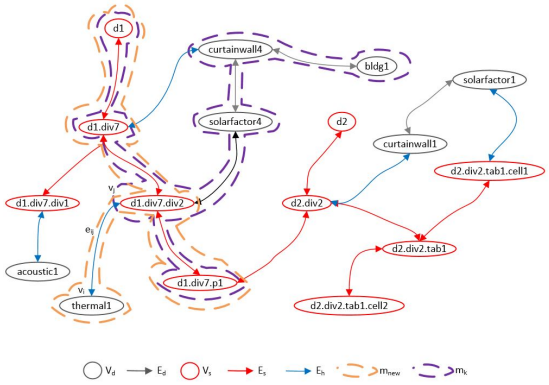
(A) Structural-based core's node of m_{new} found in existing molecule m_k.



(B) Applying the structural-based selection operator σ_s on m_k.



(C) Updated hybrid molecule m_{new}.



(D) Hybrid molecules m_{new} and m_k at termination point of Algorithm 8.

FIGURE 3.13 – Example on Algorithm 8: Appending components from each existing hybrid molecule $m_k \in M$ to a newly created molecule m_{new}.

2. **Neighboring edge e_{ij} is not hybrid** - i.e., $e_{ij} = e_s$ or $e_{ij} = e_d$ (lines 23-25): *HM_CSA* appends it to existing molecules in M_{out} where neighboring node v_j is found (line 24). Algorithm 9 details on this operation. Given the neighboring edge e_{ij} connecting current firing node v_i to neighboring node v_j, and a set of existing hybrid molecules M, the algorithm checks whether only v_j or both v_i and v_j exist in M (lines 1 and 7, respectively).

Algorithme 9 : appendToMolecules

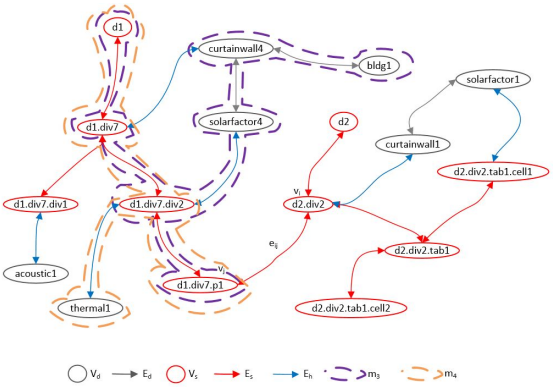
Inputs : Neighboring edge e_{ij} connecting node v_i to node v_j; Set of of existing hybrid molecules M.
Output : Updated set of hybrid molecules M with possibly new elements appended to its hybrid molecules.

```
1  if ({v_j} ∈ M AND {v_i} ∉ M) then
       // Case 1: only neighboring node v_j exists in M
2      M_j ← getMolecules(v_j) ;                    // set of molecules containing v_j, M_j ⊆ M
3      foreach m_j ∈ M_j do
4          m_j ← m_j ∪ {e_ij} ;                          // appends e_ij to matched molecule m_j
5      end
6  else
7      if ({v_j} ∈ M AND {v_i} ∈ M) then
           // Case 2: both neighboring node v_j and firing node v_i exist in M
8          M_i ← getMolecules(v_i) ;                // set of molecules containing v_i, M_i ⊆ M
9          M_j ← getMolecules(v_j) ;                // set of molecules containing v_j, M_j ⊆ M
10         if (getType(e_ij) == "Structural") then
11             m_temp ← σ_s(M_i ∪ M_j) ∪ {e_ij} ;   // merges structural-based edges and nodes of
                 hybrid molecules in M_i and M_j (by using σ_s as structural-based selection
                 operator) and appends e_ij.
12         else
13             m_temp ← σ_d(M_i ∪ M_j) ∪ {e_ij} ; // merges domain-specific edges and nodes of hybrid
                 molecules in M_i and M_j (by using σ_d as dual domain-specific selection
                 operator) and appends e_ij.
14         end
15         foreach m_i ∈ M_i do
16             m_i ← m_i ∪ m_temp ;          // updates m_i by appending new content from m_temp
17         end
18         foreach m_j ∈ M_j do
19             m_j ← m_j ∪ m_temp ;          // updates m_j by appending new content from m_temp
20         end
21     end
22 end
23 return M;
```

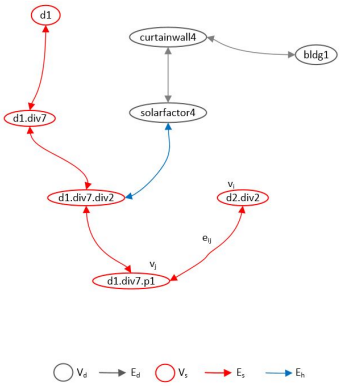
The first case remains simple as e_{ij} is appended to each hybrid molecule $m_j \in M_j$, where M_j is the set of hybrid molecules that contain v_j (lines 2-5).

For instance, Fig. 3.14a shows an explored neighboring edge e_{ij} = *(d2.div2, d1.div7.p1)* where v_j = *d1.div7.p1* is found in $m_3, m_4 \in M_j$. Thus, edge e_{ij} is simply appended to m_3 Fig. 3.14b and m_4 Fig. 3.14c.

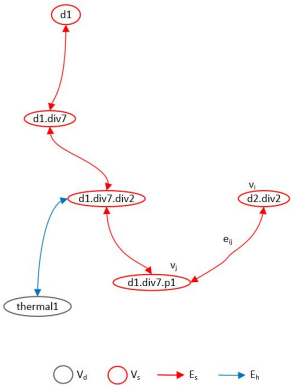
The second case (lines 8-20) is more complicated as e_{ij} should be further appended to each hybrid molecule $m_i \in M_i$, where M_i is the set of molecules involving v_i. By appending edge e_{ij} to hybrid molecules of both sets M_i and M_j, the algorithm bridges over other connected nodes among the two sets of hybrid molecules. In other words, it merges either structural-based or domain-specific components of the two sets, depending on the type of edge e_{ij} (lines 10-13). To simplify this operation, a temporary hybrid molecule m_{temp} is created involving either all structural-based or all domain-specific components of the two sets. This is done by using a structural-based or domain-specific selection operator σ_s or σ_d respectively. The edge e_{ij} is also appended to m_{temp}. After merging the appropriate parts of the two sets M_i, M_j and the edge e_{ij} in m_{temp}, the hybrid molecules of the two sets are updated with the new content of m_{temp} (lines 15-20). The final output is M, the updated set of hybrid molecules.



(A) Structural-based node v_j found in existing hybrid molecules m_3 and m_4.



(B) Updated hybrid molecule m_3 at termination point of Algorithm 9.



(C) Updated hybrid molecule m_4 at termination point of Algorithm 9.

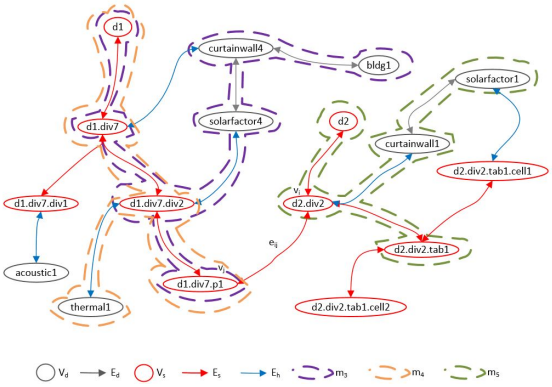
FIGURE 3.14 – Example on Algorithm 9 - Case 1: Appending neighboring edge e_{ij} to each existing hybrid molecule $m_j \in M_j$.

For instance, Fig. 3.15a shows the same example of Fig. 3.14a, however with additional hybrid molecule $m_5 \in M_i$ where $v_i = d2.div2$ is also found. Fig. 3.15b shows the constructed temporary molecule m_{temp} involving only structural-based components from $m_3 \cup m_4 \cup m_5$ since $e_{ij} = (d2.div2, d1.div7.p1)$ is a structural-based edge. Fig. 3.15c, Fig. 3.15d, Fig. 3.15e show the updated molecules m_3, m_4 and m_5 respectively, after m_{temp} has been appended to them.

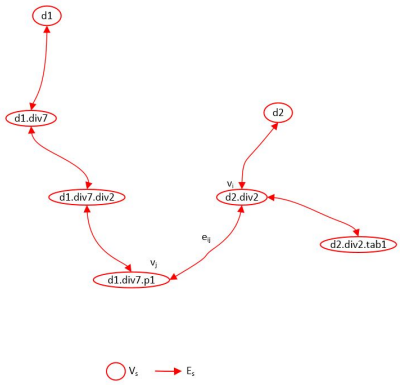
3.4.3.3 *CSA* vs *HM_CSA*

In this section, we discuss the importance of applying *HM_CSA* over a standard implementation of *CSA* using the graph $\mathcal{G}_{\delta_1}$ depicted in Fig. 3.4.

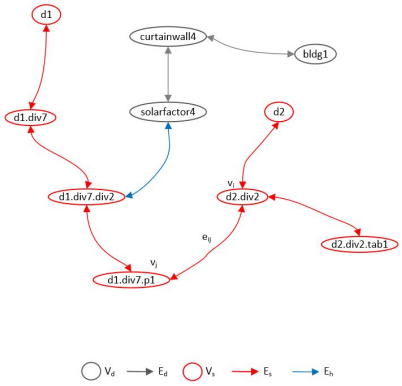
We consider the user's needs detailed in Sect. 3.1 and expressed by means of the following natural language query q = *"Solar factors of windows"*. Providing that instances of concepts *SolarFactor* and *IfcCurtainWall* are identified from q in the query interpretation module (See Sect. 3.4.2.1), the search module (See Sect. 3.4.2.2) matches nodes *solarfactor1* and *solarfactor4*, and *curtainwall1* and *curtainwall4* respectively from the graph $\mathcal{G}_{\delta_1}$, and set them as start nodes for the application of the *CSA* theory. For the sake of simplicity, we neglect post-adjustment parameters and



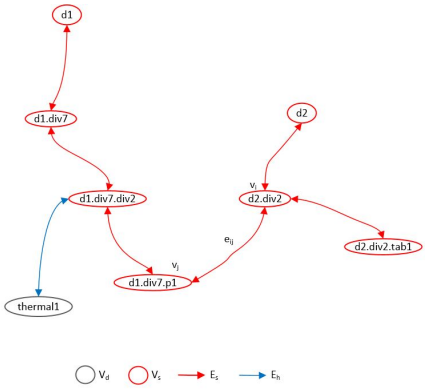
(A) Structural-based nodes v_i and v_j found in existing hybrid molecules m_5, and m_3 and m_4 respectively.



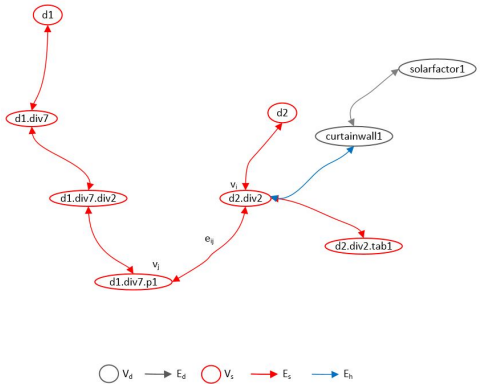
(B) Temporary hybrid molecule m_{temp} constructed after applying structural-based selection on $m_3 \cup m_4 \cup m_5$.



(C) Updated hybrid molecule m_3 at termination point of Algorithm 9.



(D) Updated hybrid molecule m_4 at termination point of Algorithm 9.



(E) Updated hybrid molecule m_5 at termination point of Algorithm 9.

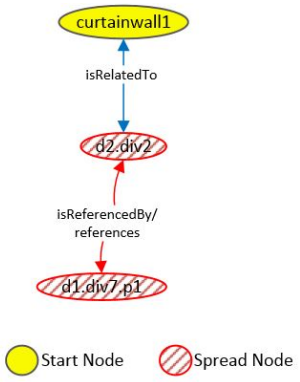
FIGURE 3.15 – Example on Algorithm 9 - Case 2: Appending neighboring edge e_{ij} to each existing hybrid molecule $m_i \in M_i$ and $m_j \in M_j$.

we choose $D = 2$ (the maximum spread distance) and $F = 0$ (the firing threshold) as pre-adjustment constraint parameters. Based on the selected start nodes, the tightly coupled semantic graph $\mathcal{G}_{\delta_1}$, and the constraint parameters, we simulate the implementation of a standard *CSA* (i.e., Algorithm 7 excluding the red-highlighted section, with V_{out} as final output) vs *HM_CSA* (i.e., Algorithm 7):

Standard *CSA*-based algorithm - The output of a standard *CSA*-based algorithm is a list of spread nodes V_{out} i.e., all nodes of the graph $\mathcal{G}_{\delta_1}$ except *acoustic1* and *d2.div2.tab1.cell2* (since their distance to start nodes is greater than $D = 2$), ranked by their activation values (See Fig. 3.16a). In some applications, the output is augmented with sub-graphs consisting of the shortest paths connecting start nodes to output nodes [81]. At termination point, one of the resulting spread nodes is *d1.div7.p1*. The latter is related to *d1* as it is part of it and to *d2* as it references one of its sections. Thus, neither presenting *d1.div7.p1* as a single node result, nor within its augmented sub-graph connecting it to the start nodes (e.g., *d1.div7.p1* → *d2.div2* → *curtainwall1*, See Fig. 3.16b) would help the user in getting this information. The latter still need to search for relevant structural and domain-specific context to understand the result, which is time and effort consuming especially in large graphs.

Spread Nodes: V_{out}	Scores: *Activation*(v_{out})
curtainwall4	*1*
solarfactor4	*1*
curtainwall1	*1*
solarfactor1	*1*
d1.div7.div2	*0.9*
d2.div2	*0.83*
d2.div7.p1	*0.76*
...	...

(A) Extract of the resulting ranked list of spread nodes.

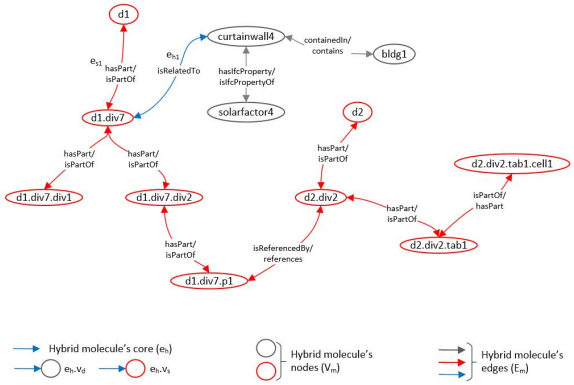


(B) Example of a sub-graph output from spread nodes to one of the initial start nodes.

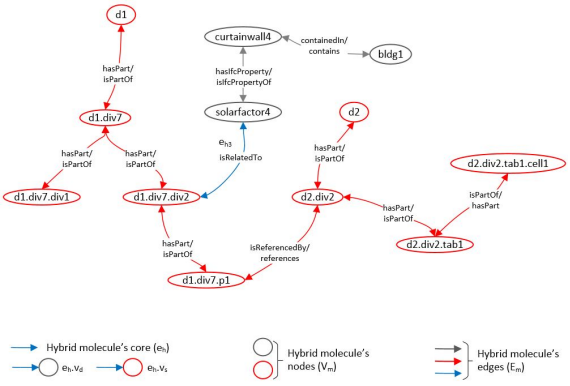
FIGURE 3.16 – Example of an output provided by *CSA* on the graph $\mathcal{G}_{\delta_1}$ of Fig. 3.4.

***HM_CSA* algorithm -** The output of *HM_CSA* consists of five hybrid molecules constructed from spread nodes and their connecting edges (See Fig. 3.17): m_1, m_3, m_4, m_5, and m_6[16] having cores e_{h1} = *(curtainwall4, d1.div7)*, e_{h3} = *(solarfactor4, d1.div7.div2)*, e_{h4} = *(thermal1, d1.div7.div2)*, e_{h5} = *(curtainwall1, d2.div2)*, and e_{h6} = *(solarfactor1, d2.div2.tab1.cell1)* respectively. The core of each molecule holds the central information from which other relevant contextual information is provided by either its connected structural nodes or its domain-specific ones. For instance, *d1.div7.p1* is now part of the five hybrid molecules. However, *d1.div7.p1* plays different roles within these molecules. In $m1$, $m3$, and m_4 *d1.div7.p1* adds containment information to e_{h1}, e_{h3}, and e_{h4} through the *isPartOf* relation. In m_5 and m_6, *d1.div7.p1* adds an inter-document link information to e_{h5} and e_{h6} through the *references* relation. This allows the user to better interpret the results, especially when he tracks fine granularity levels of the documents and cross-document dependencies between them.

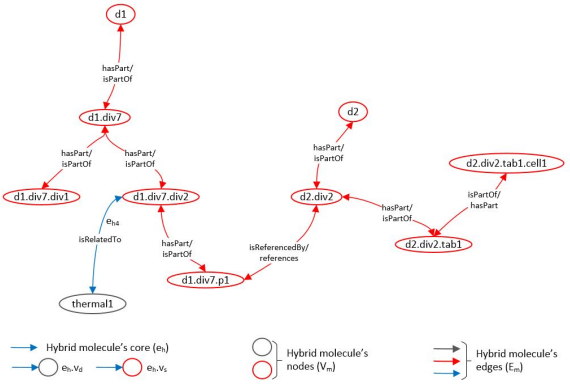
[16]For ease of presentation, the indices of the hybrid molecules follow those of their hybrid edges and are conform to the ones used in Fig. 3.6 → Fig. 3.11.



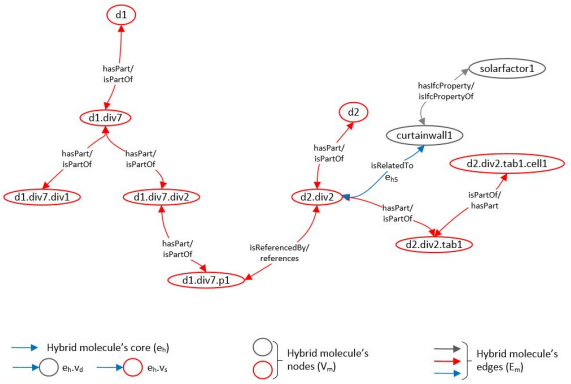
(A) Hybrid molecule m_1.



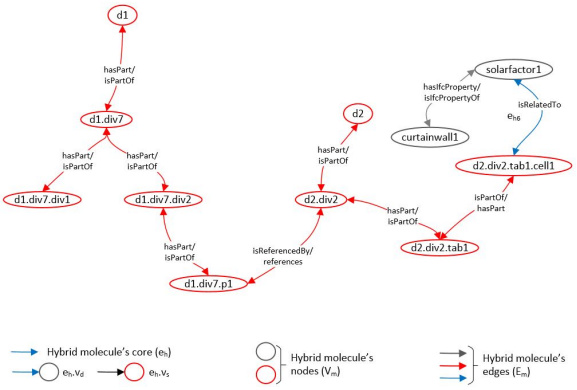
(B) Hybrid molecule m_3.



(C) Hybrid molecule m_4.



(D) Hybrid molecule m_5.



(E) Hybrid molecule m_6.

FIGURE 3.17 – Example of hybrid molecule-based output provided by *HM_CSA* on the graph $\mathcal{G}_{\delta_1}$ of Fig. 3.4.

3.4.4 Weight Mapping

In this section, we present the weighting functions used to score edges, nodes, and hybrid molecules in a tightly coupled semantic graph $\mathcal{G}_\delta$. The edges and nodes' weights are calculated during the search module, more precisely in *HM_CSA*. The hybrid molecules' weights are calculated during the Hybrid Molecule-based Ranking module (See Sect. 3.4.2.3) and correspond to the final weights of query answers.

3.4.4.1 Edge Weight Mapping

In *HM_CSA*, the edge weight mapping function f_{W_e}, used in the blue-highlighted section of Algorithm 7, is the weighting function that affects the most the output of the search module (See Sect. 3.4.2.2). This is because it directly controls the contributions of neighboring nodes on the activation value of a given node (lines 11-13), which in turn, affects the final weight of the hybrid molecule query answers encapsulating it. In the literature, however, there is no proof on the best edge weighting function. The choice remains application dependent [81]. Thus, we rely on commonly used strategies and adapt them to best suit each edge type in $\mathcal{G}_\delta$ providing rationales behind each choice of adoption or adaptation.

The weight $w_{e_{ij}}$ of an edge $e_{ij} \in E$, connecting node v_i to node v_j is calculated by $f_{W_e}(e_{ij})$ following the below strategies:

Structural-based Edge Weight - Considering that the edge e_{ij} is a structural-based edge i.e., $e_{ij} = e_s$, where e_{ij} connects two structural-based nodes v_i and v_j ($v_i, v_j \in V_s$), the corresponding edge weight $w_{e_{ij}}$ comes down to $w_{e_s} = f_{W_e}(e_s) \in [0, 1]$ and is manually acquired from the corpus expert.

The rationale is that a structural-based edge weight is a data design issue. It is set by the corpus expert to best suit the application.

For instance, consider the structural-based edge e_{s_6} = *(d1.div7.p1, d2.div2)* with $f_{Lab}(e_{s_6})$ = *"references"*, and e_{s_7} = *(d1.div7.div2, d1.div7.p1)* with $f_{Lab}(e_{s_7})$ = *"hasPart"* (See Fig. 3.4). The corpus expert may decide that e_{s_6} is more important than e_{s_7} as it provides crucial information on an inter-document link. Thus, $w_{e_{s_6}}$ may be set to 0.7 while $w_{e_{s_7}}$ may be lower and equal to 0.5

Domain-specific Edge Weight - Considering that the edge e_{ij} is a domain-specific edge i.e., $e_{ij} = e_d$, where e_{ij} connects two domain-specific nodes v_i and v_j ($v_i, v_j \in V_d$), the corresponding edge weight $w_{e_{ij}}$ comes down to $w_{e_d} = f_{W_e}(e_d)$, such that:

$$f_{W_e}(e_d) = \frac{1}{fan - in_{lab}(v_j)} \qquad \in]0, 1] \tag{3.2}$$

Where,

- $\frac{1}{fan - in_{lab}(v_j)}$ is the specificity measure of the edge e_d incoming towards v_j
- $fan - in_{lab}(v_j)$ is the number of incoming nodes towards v_j having the same label lab of e_d i.e., $lab = f_{Lab}(e_d)$

The rationale is that the specificity measure reflects the importance of a domain-specific edge w.r.t. a target node. The less incoming edges with the same label, the more important the edges become for a node. This measure is commonly used in the literature of semantic graphs (e.g., [22, 81]).

For instance, consider the domain-specific edge e_{d_2} = *(solarfactor4, curtainwall4)* with $lab_2 = f_{Lab}(e_{d_2})$ = *"isIfcPropertyOf"* (See Fig. 3.4). $fan - in_{lab_2}(e_{d_2}) = 1$ as it is the only domain-specific edge with label lab_2 incoming towards *curtainwall4*, thus $w_{e_{d_2}} = \frac{1}{1} = 1$.

Hybrid Edge Weight - Considering that the edge e_{ij} is a hybrid edge i.e., $e_{ij} = e_h$, where e_{ij} connects a domain-specific node $v_i \in V_d$ to a structural-based node $v_j \in V_s$ or vice versa, the corresponding edge weight $w_{e_{ij}}$ comes down to $w_{e_h} = f_{W_e}(e_h)$, such that:

$$f_{W_e}(e_h) = TF(s_i, v_j) \times IDF(s_i, V_s) \qquad \in [0, 1[\tag{3.3}$$

Where,

- $TF(s_i, v_j) \times IDF(s_i, V_s)$ is inspired by the notion of TF-IDF (Term Frequency - Inverse Document Frequency) score;
- $TF(s_i, v_j)$ is the frequency of the value s_i of the domain-specific node $v_i \in V_d$ occurring in a structural-based node $v_j \in V_s$ (e.g., a document or a part of a document), where $s_i = f_{Val}(v_i)$. It is calculated as follows:

$$TF(s_i, v_j) = \frac{NbOcc(s_i, v_j)}{Max(NbOcc(s_k, v_j))} \qquad \in]0, 1] \tag{3.4}$$

  Such that,

  - $NbOcc(s_i, v_j)$ is the number of occurrences of the string value of v_i in the content of its related structural-based node v_j.

    Note that, the minimum value of $NbOcc(s_i, v_j)$ is equal to 1 since the presence of the hybrid relation means that, during the annotation phase, an occurrence of s_i was identified in v_j and lead to the creation of the hybrid edge (as discussed in Chapter 2, in Sect. 2.3.3.3).
  - $Max(NbOcc(s_k, v_j))$ is the maximum number of occurrences of the string value s_k of any domain-specific node $v_k \in V_d$ connected to the same structural-based node v_j through another hybrid edge e_{kj}. Thus, $TF(s_i, v_j)$ is never equal to 0.

- $IDF(s_i, V_s)$ is the inverse frequency of the value s_i of the domain-specific node $v_i \in V_d$ occurring in the set of all structural-based nodes V_s. It is calculated as follows:

$$IDF(s_i, V_s) = 1 - \frac{DF(s_i, V_s)}{|V_s|} \qquad \in [0, 1[\tag{3.5}$$

  Such that,

- $DF(s_i, V_s)$ is the number of structural-based nodes v_s involving at least one occurrence of the string value s_i of v_i.

  Note that the minimum value of $DF(s_i, V_s)$ is equal to 1 since s_i has at least occurred in one structural-based node $v_s \in V_s$, specifically in $v_j \in V_s$. Thus, $IDF(s_i, V_s)$ is never equal to 1.
- $|V_s|$ is the total number of structural-based nodes in the graph $\mathcal{G}_\delta$.

The rationale is that, in $\mathcal{G}_\delta$, a domain-specific node v_i describes the content of a structural-based node v_s, thus the value of a v_i can be perceived as a term and v_s as a document in a *TF-IDF* like notion (in a similar vein with other IR applications, such as [95]). Consequently, a hybrid edge weight w_{e_h} is directly proportional to the number of occurrences of the domain-specific information in a structural-based element and inversely proportional to the number of occurrences of the domain-specific information contained in other structural-based elements.

For instance, consider the hybrid edge e_{h_5} = *(curtainwall1, d2.div2)* (See Fig. 3.4) with $f_{Lab}(e_{h_5})$ = *"isRelatedTo"*. Given v_i = *curtainwall1* and has a string value $s_i = f_{Val}(v_i)$ = *'window frames"*, and v_j = *d2.div2* and has a string content $s_j = f_{Val}(v_j)$ = *"Characteristics of Window Frames"* $\Rightarrow TF(s_i, v_j) = \frac{1}{1} = 1$ since $NbOcc(s_i, v_j) = 1$ and there is no other value of a domain-specific node v_k connected to v_j that occurs several times in its content s_j. Also, considering that s_i did not occur in any other structural-based node $v_s \in V_s$ and $|V_s| = 10$, $DF(s_i, V_s) = 1 \Rightarrow IDF(s_i, V_s) = 1 - \frac{1}{10} \simeq 0.9$. Thus, $w_{e_{h_4}} = 1 \times 0.9 = 0.9$

3.4.4.2 Node Weight Mapping

In any *CSA*-based application, the node weights come down to their final activation values [24]. Naturally, the activation function reflects the contribution of neighboring nodes considering the strength of linking edges (as described in Sect. 3.4.3.2). Along the same lines, we consider that the node weights w_{v_s} of a structural-based node $v_s \in V$ and w_{v_d} of a domain-specific node $v_d \in V$ are calculated by $f_{W_v}(v)$ based on the final activation value of a node $v \in V$ regardless of its type i.e,:

$$f_{W_v}(v) = Activation(v) \quad \in [0, 1] \quad , \forall v \in V. \tag{3.6}$$

The rationale is that given the different strategies that are applied to calculate the edge weights (See Sect. 3.4.4.1) based on their types, the activation function implicitly weighs the nodes differently considering their nature i.e., whether they are structural-based nodes or domain-specific nodes.

For instance, consider the structural-based node v_j = *(d2.div2)* in Fig. 3.18. The activation value $Activation(v_j)$ is the result of the contributions of firing nodes V_i = *(curtainwall1, d2.div2.tab1, d1.div7.p1, d2)* at different points of time t_k of Algorithm 7. At t_0, consider that the domain-specific node v_i = *(curtainwall1)*, where $v_i \in V_i$,

spreads out its activation to its neighboring node v_j. Providing that the initial activation values of nodes v_i and v_j are set to 1 and 0 respectively i.e., $Activation(v_i) = 1$ and $Activation(v_j) = 0$, and $\gamma = 0.2$ i.e., the decay factor $(1 - \gamma)$ is equal to 0.8: $\Delta_{input}(v_j) = 1 \times 0.9 \times 0.8 = 0.72 \Rightarrow Input(v_j) = 0 + 0.72 = 0.72 \Rightarrow Output(v_j) = Activation(v_j) = 0.72$. Similarly, at t_k, another node $v_i \in V_i$ spreads out its activation to v_j, until the termination point t_{END}, where the final activation value of v_j is set to the contributions of all nodes $v_i \in V_i$. Consequently, if at t_{END} of Algorithm 7 $Activation(v_j) = 1 \Rightarrow$ the node weight w_{v_j} of the structural-based node *d2.div2* comes down to $w_{v_j} = 1$. This means that, over spread iterations, firing nodes V_i strengthened the importance of the neighboring node $v_j = (d2.div2)$ as it is possibly relevant w.r.t. initial start nodes.

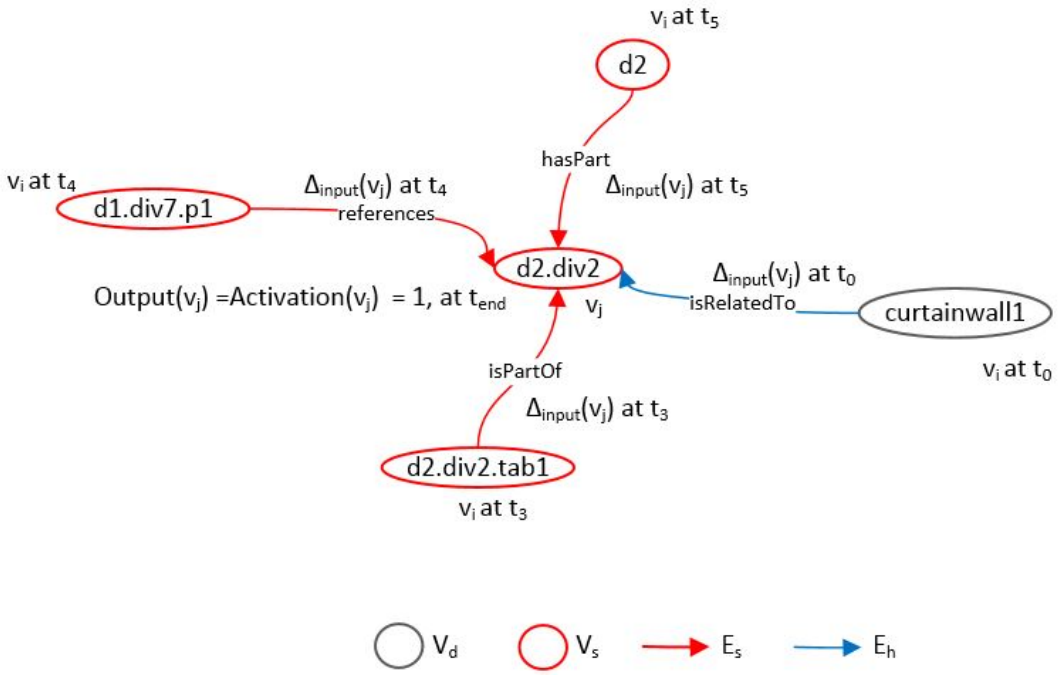


FIGURE 3.18 – Contributions of firing nodes in the calculation of the weight of a neighboring node over spread iterations at different points of time.

3.4.4.3 Hybrid Molecule Weight Mapping

A *Hybrid Molecule* is a novel data structure that we propose for query answers. Thus, we propose a novel weighting function associated to it. The weight w_m of a hybrid molecule $m \in M$ is calculated by $f_{W_m}(m)$, used in the *HM_Score* function of Algorithm 5 (See Sect. 3.4.2.3), based on the weights of its nodes:

$$f_{W_m}(m) = \frac{\alpha \times \sum w_{v_s} + \beta \times \sum w_{v_d}}{|V_m|} \times w_{e_h} \qquad \in [0,1] \tag{3.7}$$

Where,

- w_{v_s} and w_{v_d} are the weights computed by $f_{W_v}(v)$ and assigned to a structural-based node $v_s \in V_m$ and a domain-specific node $v_d \in V_m$ respectively.
- $|V_m|$ is the total number of nodes in m.
- α and β are the weight parameters that balance the contribution of the structural and domain-specific parts of m, such that α and $\beta \in [0, 1]$.
- w_{e_h} is the hybrid edge weight of the core e_h of m

  Note that, α and β come down to *rank_params* in Algorithm 5.

The rationale is that, as the hybrid molecules are constructed progressively, they are changing from one spread iteration to the other. Hence, their weights are the best to be calculated based on the contributions of their components at termination point of *HM_CSA*. This is done using the final activation values of the nodes composing them which implicitly reflect the contributions of the edge components as well. For the sake of simplicity, this is expressed through an equation of weighted average values of structural-based and domain-specific contributions within a hybrid molecule, which is then multiplied by the weight of the hybrid edge identifier of the molecule. The latter is done for two main reasons: (1) the hybrid edge is the unique identifier of a molecule as it holds core information (as discussed in Sect. 3.3.2), (2) some molecules may involve the same structural-based and domain-specific components yet different core information, thus the hybrid edges should leverage their weights so they could be scored differently and conveniently.

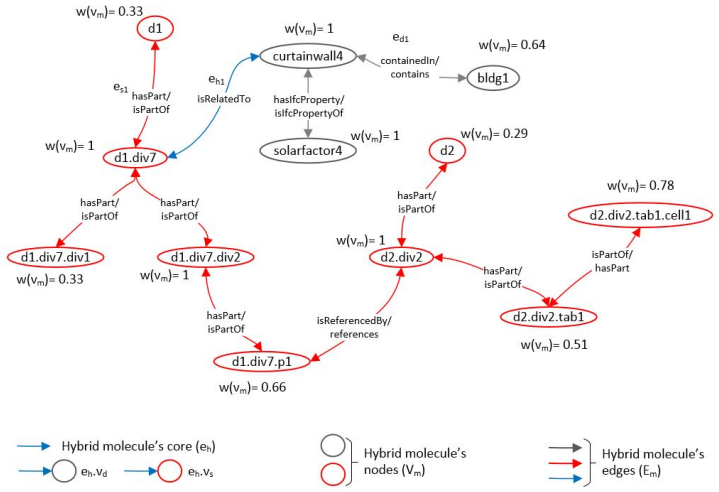
For instance, Fig. 3.19 shows three hybrid molecules m_1, m_3 and m_4 together with their component nodes' weights from the output of *HM_CSA* applied on the example described in Sect. 3.4.3.3. Considering $\alpha = \beta = 1$:

$$w_{m_1} = \frac{1 \times (0.33 + 0.33 + 0.83 + 1 + 0.29 + 1 + 0.66 + 0.78 + 0.51) + 1 \times (1 + 1 + 0.64)}{12} \times 0.9 = 0.63$$

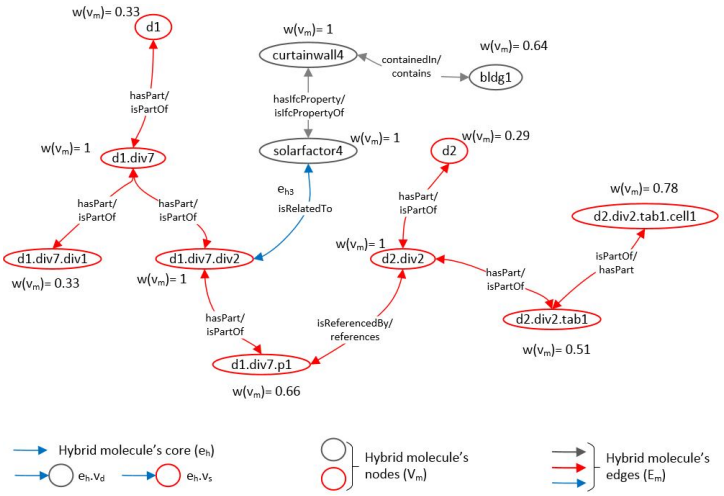
$$w_{m_3} = \frac{1 \times (0.33 + 0.33 + 0.83 + 1 + 0.29 + 1 + 0.66 + 0.78 + 0.51) + 1 \times (1 + 1 + 0.64)}{12} \times 0.8 = 0.56$$

$$w_{m_4} = \frac{1 \times (0.33 + 0.33 + 0.83 + 1 + 0.29 + 1 + 0.66 + 0.78 + 0.51) + 1 \times (0.46)}{10} \times 0.9 = 0.56$$

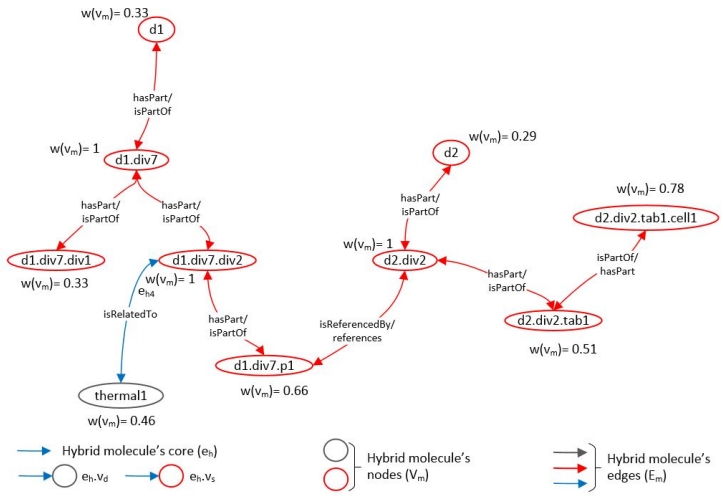
Although m_1 and m_3 have exactly the same structural-based and domain-specific components, their hybrid edges e_{h_1} and e_{h_3} respectively are different and have different weights. $w_{e_{h_1}} = 0.9$, being higher than $w_{e_{h_3}} = 0.8$, the hybrid molecule weight w_{m_1} of m_1 is higher than the weight w_{m_3} of m_3.



(A) Hybrid molecule m_1.



(B) Hybrid molecule m_3.



(C) Hybrid molecule m_4.

FIGURE 3.19 – Hybrid molecules m_1, m_3 and m_4 together with their weighted components at termination point of Algorithm 7.

Considering $\alpha = 1, \beta = 0.5$ to emphasize the contribution of structural-based nodes on the weights of m_1 and m_4:

$$w_{m_1} = \frac{1 \times (0.33 + 0.33 + 0.83 + 1 + 0.29 + 1 + 0.66 + 0.78 + 0.51) + 0.5 \times (1 + 1 + 0.64)}{12} \times 0.9 = 0.52$$

$$w_{m_4} = \frac{1 \times (0.33 + 0.33 + 0.83 + 1 + 0.29 + 1 + 0.66 + 0.78 + 0.51) + 0.5 \times (0.46)}{10} \times 0.8 = 0.54$$

Since m_4 involves relatively less domain-specific nodes compared to m_1, it has been less affected when reducing the contribution of domain-specific nodes on the weights of the hybrid molecules.

3.5 Summary

This chapter tackles the problem of handling IR over a *tightly coupled semantic graph* representing a heterogeneous document corpus. The main purpose is to provide the users with augmented search results i.e., query answers including both the structural and the domain-specific dimensions of the document corpus. This is not covered by current SIR systems which mainly neglect, in their search results, relevant granularity levels of the documents and dependencies between them. The contributions of this chapter come down to: (i) formally defining *Hybrid Molecules*, a novel data structure for query answers based on a tightly coupled semantic graph's definition and the definitions of molecules in the literature, (ii) constructing the *Hybrid Molecules* progressively throughout a *Hybrid Molecule-based Query Processing* which involves *Query Interpretation*, *Hybrid Molecule-based Search*, *Hybrid-Molecule-based Ranking*, and *Hybrid Molecule-based Presentation*, (iii) providing a novel graph-based search algorithm *HM_CSA* inspired by the *CSA* theory, which details on the search module, and (iv) *Weight Mapping* functions that provide edge, node and hybrid molecule weight strategies handling the characteristics of a tightly coupled semantic graph and used in the ranking module. To our knowledge, the *Hybrid Molecules* are the first query answers providing the user with (i) relevant granularity levels of the documents, (ii) relevant inter and intra-document dependencies, and (iii) contextual information helping the users in interpreting the search results and tracking cross document dependencies. Several experiments were conducted to validate our proposal within real-world applications. These experiments are dedicated to Chapter 4.

Chapter 4

Experimental Evaluation

"I did not think; I experimented."
- Wilhelm Conrad Röntgen

In Chapter 2, we propose a collective knowledge representation of a heterogeneous document corpus, which comes down to a tightly coupled semantic graph generated based on the infrastructure provided by a novel multi-layered ontology, entitled *LinkedMDR*. In Chapter 3, we propose an innovative IR model which exploits such a graph in a query processing pipeline relying on a novel data structure for query answers, the Hybrid Molecules, and a set of dedicated algorithms. **This chapter addresses the validation of these proposals in the context of a domain-specific application, particularly the AEC industry.** We implemented a Java-based prototype made of two main modules: an *LMDR Annotator* and an *HM Query Processor*. The former automatically annotates a heterogeneous document corpus and generates the tightly coupled graph. The latter exploits the generated semantic graph and provides a ranked list of Hybrid Molecule-based answers in response to a user's natural language query. The two modules provide a means to evaluate annotations of the heterogeneous document corpus and query results separately. The experiments were conducted on construction projects provided by Nobatek/INEF4 in order to demonstrate that the proposed contributions are applicable in real-world applications.

4.1 Introduction

This chapter describes the experimental study we conducted in order to validate the contributions of Chapter 2 and Chapter 3 in the context of the AEC industry, using real-world construction projects from Nobatek/INEF4. We set two main global objectives: (i) evaluate the knowledge representation of a heterogeneous document corpus on the basis of the contributions of Chapter 2 (*Objective 1*), and (ii) evaluate the quality of the IR model on the basis of contributions of Chapter 3. To so do, we first implemented *FEED2SEARCH* framework layers (See Chapter 1) within a Java-based prototype as follows:

- We used *LinkedMDR* OWL file (lmdr.owl)[1] for the *Knowledge Representation Layer*. This version is built on OWL 2 in the Protégé environment and serialized in RDF/XML. It relies on DC [29], TEI [98] and MPEG-7 [97] standards for the Standardized Metadata Layer and on an adapted ifcOWL [18]-based ontology as an external domain-specific ontology for the Pluggable Domain-Specific Layer;

- We developed a module for the *Indexing* Layer, entitled *LMDR Annotator*, in order to generate a tightly coupled semantic graph representing a heterogeneous document corpus. Although *LMDR Annotator* provides automatic pipelines from the stage of upload of the document corpus to the stage of the generation of the corresponding semantic graph, the users still have the possibility to intervene in all the intermediate stages in order to collect the generated results for experimental purposes;

- We developed a module for the *Query processing* Layer, entitled *HM Query Processor*, in order to generate a ranked list of Hybrid Molecule-based query answers w.r.t. a user's natural language query. Although the main purpose of the *Query Processing* layer of *FEED2SEARCH* is to output the hybrid molecules in SERP-like results, we do not so far focus on GUI issues while implementing *HM Query Processor* as it is not the main contribution of the thesis. Instead, we give an intermediary visualization layout of the hybrid molecules, which, with our assistance, is presented to users in the AEC industry for experimental evaluations.

The remainder of this chapter is organized as follows. Sect. 4.2 and Sect. 4.3 describe technical details regarding the so far implemented sub-modules of *LMDR Annotator* and *HM Query Processor* respectively. Sect. 4.4 and Sect. 4.5 present experimental protocols and results of two different sets of experiments in the context of knowledge representation (for *Objective 1*) and IR (for *Objective 2*) respectively. Sect. 4.6 concludes the chapter.

[1] Available at `http://spider.sigappfr.org/linkedmdr/`

4.2 LMDR Annotator

We developed *LMDR Annotator*, which stands for ***L***inked ***MDR Annotator***, within *FEED2SEARCH* Java-based prototype. Its main purpose is to automatically annotate a given document corpus δ. It generates a tightly coupled semantic graph $\mathcal{G}_\delta$ representing δ. *LMDR Annotator* has several sub-modules involving services of the Indexing layer of FEED2SEARCH (See Fig. 1.5 in Chapter 1). It is, in fact, an implementation of the tight coupling algorithm that we presented in Sect. 2.3.3.3 of Chapter 2 (See Algorithm 1). In this section, we first present an overview of the different sub-modules of *LMDR Annotator* (See Sect. 4.2.1), then we provide implementation details regarding the underlying technologies, APIs, and web services adopted for each of its sub-modules (See Sect. 4.2.2, 4.2.3 4.2.4, 4.2.5, 4.2.6).

4.2.1 Overall Architecture

Fig. 4.1 illustrates an overview of the different sub-modules of *LMDR annotator*, which is based on *LinkedMDR* ontology implemented in the context of the AEC industry. The user's input is the heterogeneous document corpus δ and the system's final output is the graph $\mathcal{G}_\delta$. We choose RDF for the serialization of $\mathcal{G}_\delta$ ($\mathcal{G}_\delta.rdf$) as it is a widely used and reliable data model for representing a semantic and extensible graph [104].

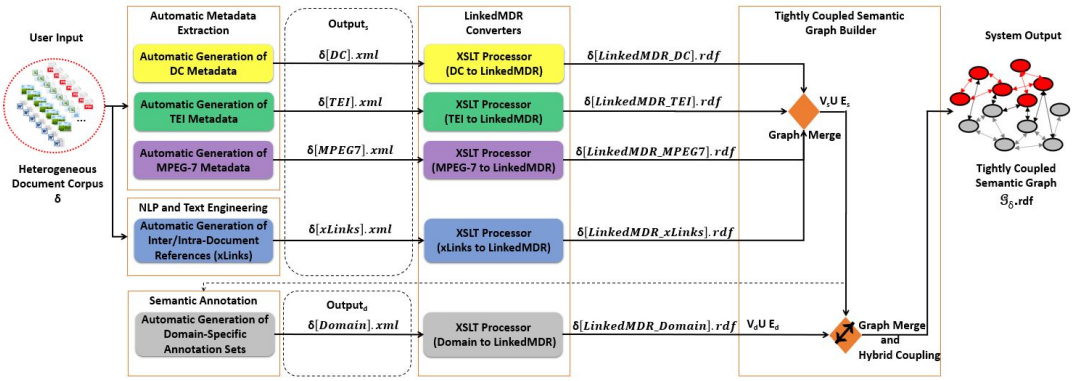


FIGURE 4.1 – Overview of the different sub-modules of LMDR Annotator.

LMDR Annotator is made of five main sub-modules: Automatic Metadata Extraction, NLP and Text Engineering, Semantic Annotation, *LinkedMDR* converters, and Tightly Coupled Semantic Graph Builder. The first two sub-modules output structural-based descriptors (i.e., $Output_s$ in Algorithm 1) in an XML format. The semantic Annotation sub-module outputs domain-specific descriptors (i.e., $Output_d$ in Algorithm 1) in an XML format. The *LinkedMDR* Converters sub-module transforms XML outputs generated by the three previous sub-modules into RDF format following semantics of *LinkedMDR* ontology. The tightly coupled semantic graph builder sub-module progressively merges the RDF outputs to sequentially generate nodes and edges of the tightly coupled semantic graph $\mathcal{G}_\delta$. It also ensures the hybrid

coupling, which creates hybrid edges between structural-based nodes and domain-specific ones.

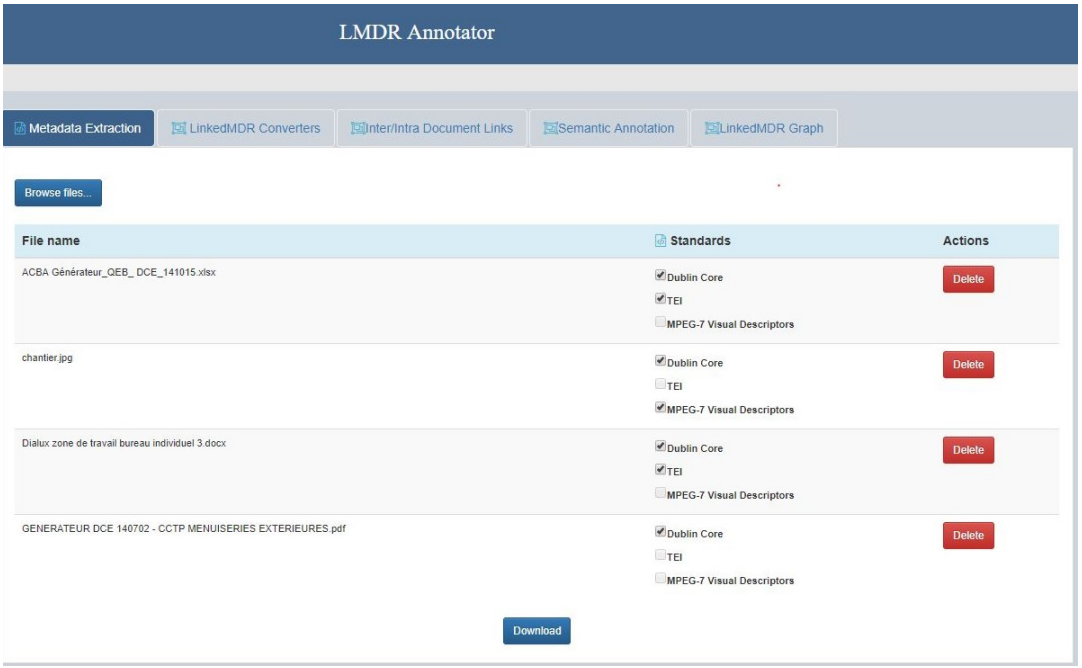


FIGURE 4.2 – Screen-shot of the current version of LMDR Annotator.

Fig. 4.2 shows a screen-shot of the implemented prototype. For instance, it provides an example of the Automatic Metadata Extraction module where the user uploads a set of heterogeneous documents simultaneously (e.g., an XLSX technical specification document, an on-site JPG image, a PDF document describing the exterior carpentry, etc.). The system automatically checks, for each document, the sub-modules capable of generating descriptors based on existing standards. However, the user still have the possibility to uncheck them. When he clicks on the *Download* button, the system automatically downloads XML files describing these documents based on the chosen standards. One document could have more than one XML file describing it. For instance, the first XLSX will have two XML documents, one involving DC descriptors and another one involving TEI descriptors.

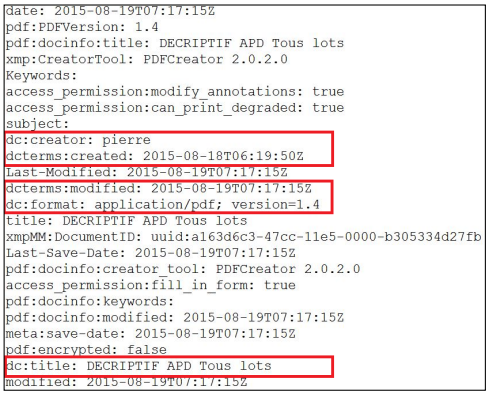
4.2.2 Automatic Metadata Extraction

The main purpose of this sub-module is to automatically generate descriptors of DC [29], TEI [98] and MPEG-7 [97] that are reused in *LinkedMDR* ontology based on existing tools, APIs or web services:

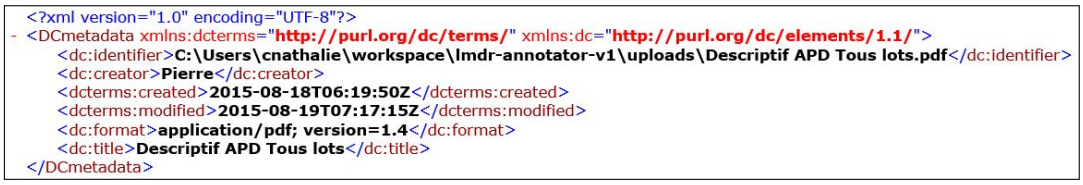
4.2.2.1 Automatic Generation of DC Metadata

Among the various DC-based tools[2], *LMDR Annotator* uses the Apache Tika 1.4 Toolkit[3] as it provides a fully automatic technique to generate, from a wide variety of document formats(JPG, TXT, PDF, etc.), a list of DC descriptors covering all the elements reused in *LinkedMDR*.

The Apache Tika toolkit automatically detects and extracts metadata and text using metadata harvesting techniques [42] and basic textual content extraction respectively. Metadata harvesting techniques detect META tags either produced by humans or supported by the Software tool at creation or update time of the document [73]. Although these META tags follow several descriptors, we only extract DC compatible metadata [29] (e.g., *dc:creator*, *dc:format*, *dc:terms:created*). As shown in Fig. 4.3a, the output generated by this API is in the form of plain text or XHTML. Thus, *LMDR Annotator* transforms the selected DC metadata descriptors into an XML format following specifications of DC standard i.e., in DC XML encoding format[4] (See Fig. 4.3b).



(A) Extract of Tika API output in form of plain text.



(B) Transformed Tika output in DC XML.

FIGURE 4.3 – Example of automatic generation of DC metadata on a PDF document describing general technical specifications.

[2]E.g., the Dublin Core Advanced Generator available at `https://nsteffel.github.io/dublin_core_generator/generator.html`, the Editor-Converter Dublin Core available at `http://library.kr.ua/dc/dcedituni`

[3]Available at `https://tika.apache.org/`

[4]Guidelines available at `http://dublincore.org/documents/dc-xml-guidelines/`

4.2.2.2 Automatic Generation of TEI Metadata

Among few available TEI-based tools[5], *LMDR Annotator* uses Oxgarage[6] web RESTful service as it provides a fully automatic technique to generate, from a wide variety of document formats (e.g., TXT, DOCX, XLSX, etc.), a reliable list of TEI [98] descriptors covering all the elements reused in *LinkedMDR*. It is also recommended by the TEI consortium[7].

Oxgarage web service is based on the Enrich Garage Engine[8] which provides services for data conversion, transformation and validation. Among the underlying converters, the TEI converter is based on eXstensible Stylesheets Language (XSL) to convert between different forms of XML documents. The output of the web service calling the TEI converter is an XML format following specifications of TEI standard. Fig. 4.4 shows an extract of the XML output of Oxgarage web service which generates TEI structural metadata describing the structural decomposition of a word document regardless its language (e.g., French) into a section containing a list and a paragraph.

```
<div>
    <head>
        <hi rend="italic">Au début du chantier</hi>
    </head>
    <p>Dans le délai fixé au calendrier prévisionnel des travaux,
    l'Entrepreneur du présent lot devra prévoir la présentation
    de pré-prototypes et d'échantillons selon le processus suivant :
    </p>
    <list type="unordered">
        <item>Diffusion au Maître d'Œuvree au bureau de contrôle des plans d'exécution.</item>
        <item>Présentation au Maître d'Œuvre de pré-prototypes (ouvrages partiels dont tous
        les composants sont facilement démontables), permettant l'examen des différents éléments
        dans leur ordre de montage, et l'analyse critique des points importants.</item>
        <item>Présentation d'échantillons complémentaires relatifs au procédé.</item>
        <item>Pièces de fixation et joints d'étanchéité avec les ouvrages adjacents.</item>
    </list>
    <p>Tous ces échantillons seront fixés sur un panneau présentoir et resteront à demeure
    sur le chantier, dans le local prévu à cet effet, jusqu'à la réception des travaux.
    </p>
</div>
```

FIGURE 4.4 – Extract of the XML output of Oxgarage web service for the automatic generation of TEI metadata on a WORD document describing technical specifications regarding the Exterior Carpentry.

4.2.2.3 Automatic Generation of MPEG-7 Metadata

In *LinkedMDR* ontology, we so far reused MPEG-7 descriptors related to the image metadata with different levels of precision (e.g., $\langle mpeg7 : Image \rangle$, $\langle mpeg7 : StillRegion \rangle$), its related visual descriptors, and semantic descriptors. For practical reasons, in the implementation of *LMDR Annotator*, we only focus on the automatic generation of MPEG-7 visual descriptors for two main reasons: (i) the automatic

[5] E.g., GeneRation Of BIbliographic Data (GROBID) available at `https://grobid.readthedocs.io/en/latest/`

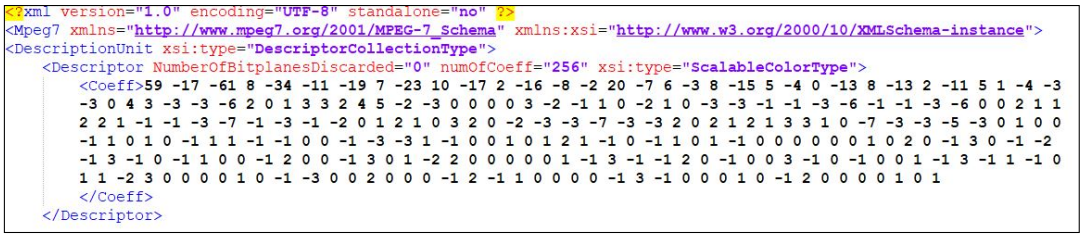
[6] Related API available at `https://github.com/sebastianrahtz/oxgarage`

[7] Recommendations available at http://www.tei-c.org/tools/

[8] Developed by Poznan Super computing and Networking Center, and Oxford University Computing Services for the EU-funded ENRICH project, `https://sourceforge.net/projects/enrich-ege/`

generation of the different granularity levels of the image metadata requires computer vision and machine learning algorithms [71], which is not the core purpose of the thesis, and (ii) the automatic generation of semantic descriptors still require human intervention.

Among several MPEG-7-based tools[9], *LMDR Annotator* uses MPEG-7 Visual Descriptors[10] as it provides a fully automatic technique to generate, from image content files (e.g., JPG, PNG, etc.), MPEG-7 visual descriptors [97] reused in *LinkedMDR*. The library is the implementation of the Open Source Java Content Based Image Retrieval Library (Lire) [61] in c#. It generates information regarding *Scalable Color*, *Color Layout*, *Dominant Colors*, and *Edge Histogram*. *LMDR Annotator* calls the library through command line executed in Java and outputs an XML format following specifications of MPEG-7 standard [97] (See Fig. 4.5).



```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
<Mpeg7 xmlns="http://www.mpeg7.org/2001/MPEG-7_Schema" xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance">
<DescriptionUnit xsi:type="DescriptorCollectionType">
    <Descriptor NumberOfBitplanesDiscarded="0" numOfCoeff="256" xsi:type="ScalableColorType">
        <Coeff>59 -17 -61 8 -34 -11 -19 7 -23 10 -17 2 -16 -8 -2 20 -7 6 -3 8 -15 5 -4 0 -13 8 -13 2 -11 5 1 -4 -3
        -3 0 4 3 -3 -3 -6 2 0 1 3 3 2 4 5 -2 -3 0 0 0 0 3 -2 -1 1 0 -2 1 0 -3 -3 -1 -1 -3 -6 -1 -1 -3 -6 0 0 2 1 1
        2 2 1 -1 -1 -3 -7 -1 -3 -1 -2 0 1 2 1 0 3 2 0 -2 -3 -3 -7 -3 -3 2 0 2 1 2 1 3 3 1 0 -7 -3 -3 -5 -3 0 1 0 0
        -1 1 0 1 0 -1 1 1 -1 -1 0 0 -1 -3 -3 1 -1 0 0 1 0 1 2 1 -1 0 -1 1 0 1 -1 0 0 0 0 0 0 1 0 2 0 -1 3 0 -1 -2
        -1 3 -1 0 -1 1 0 0 -1 2 0 0 -1 3 0 1 -2 2 0 0 0 0 0 1 -1 3 -1 -1 2 0 -1 0 0 3 -1 0 -1 0 0 1 -1 3 -1 1 -1 0
        1 1 -2 3 0 0 0 0 1 0 -1 -3 0 0 2 0 0 0 -1 2 -1 1 0 0 0 0 -1 3 -1 0 0 0 1 0 -1 2 0 0 0 0 1 0 1
        </Coeff>
    </Descriptor>
```


FIGURE 4.5 – Extract of the XML output of MPEG-7 Visual Descriptors library for the automatic generation of MPEG-7 metadata on a JPG photo capturing on-site construction works.

4.2.3 NLP and Text Engineering for Automatic Generation of Inter/Intra-Document References

We focus on the automatic generation of specific types of inter and intra-document dependencies which are cross-references between documents or parts of documents identified from their textual content. These relations are agreed to be frequently encountered by actors of the construction projects and very important in tracking complementary information among several related documents.

We use the General Architecture for Text Engineering (GATE) API[11] as it provides robust techniques for NLP and Text Engineering [26] to generate, from textual content of various document formats (e.g., TXT, PDF, DOCX, etc.), annotation sets based on predefined rules and knowledge resources. Consequently, we consider cross-references as annotation sets identified from the occurrences of pre-defined expressions in the heterogeneous document corpus. For instance, *"cf."*, *"voir"*, *"performances selon"*, *"se référer"* are examples of frequently encountered expressions in

[9] E.g., Caliph & Emir Java Open source Library available at `http://www.semanticmetadata.net/features/`, MPEG-7 Feature Extraction Library available at `http://www.cs.bilkent.edu.tr/~bilmdg/bilvideo-7/Software.html`

[10] Available at `https://chatzichristofis.info/?page_id=19`

[11] Gate Developer 8.1 available at `https://gate.ac.uk/download/`

French construction related documents. In order to generate the required annotation sets, *LMDR Annotator* executes a pre-configured GATE pipeline application and provides it with the required knowledge resources. The GATE pipeline is made of the following main processing resources:

- *Document Reset*, *Tokeniser*, *Sentence Splitter*, *POS Tagger*, and *Morphological Analyser* for linguistic pre-processing of the textual content;
- *ANNIE Gazetteer* which consists of a pre-defined list of expressions that yield inter and intra-document references;
- *OntoGazetter* which creates mappings between lists containing documents' information (e.g., lists of document names dynamically created at execution time of *LMDR Annotator*) and corresponding *LinkedMDR* instances[12] (e.g., a mapping between a document name and its corresponding *LinkedMDR* instance URI);
- *JAPE Transducer*[13] which allows to recognize regular expressions and to subsequently create inter and intra-document references.

For instance, Fig. 4.6 shows an example of a pre-defined rule in the *JAPE Transducer*, which detects an inter-document reference *InterDocLink* whenever it identifies a *Link* entity (i.e., an entry from the pre-defined list of expressions in the *ANNIE Gazetteer*), followed by 0 to 3 *Token* entities (i.e., entries from the pre-processed list of tokens), followed by a *Document* entity (i.e., an entry from the generated list of document information in the *OntoGazetteer*). *LMDR Annotator* outputs the generated annotation sets in XML format. An extract of the generated output following the execution of the above *JAPE* rule is illustrated in Fig. 4.7 where *InterDocLink* represents an annotation set identified from the following French textual content: *"Pour plus de détails, cf. notice thermique"*.

```
Phase: process2
Input: Document Link Token
Options: control=brill debug=true

Rule: InterDocLinkType1
(
  ({Link}) :xLink
  ({Token}) [0,3]
  ({Document}) :doc
):link1
--> :link1.InterDocLink = {Rule="InterDocLink1",docname=:doc@string, linkname=:xLink@string}
```

FIGURE 4.6 – Example of a pre-define rule in a *JAPE Transducer*.

[12]Instances of documents are created by any of the *LinkedMDR* converters' sub-module (e.g., DC to LinkedMDR).

[13]JAPE is a Java Annotation Patterns Engine providing finite state transduction over annotations based on regular expressions.

```
<Annotation Id="1468" Type="InterDocLink" StartNode="102" EndNode="122">
<Feature>
  <Name className="java.lang.String">linkname</Name>
  <Value className="java.lang.String">cf.</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">docname</Name>
  <Value className="java.lang.String">notice thermique</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">Rule</Name>
  <Value className="java.lang.String">InterDocLink1</Value>
</Feature>
</Annotation>
```

FIGURE 4.7 – Extract of the generated annotation sets describing inter document references using the GATE API.

4.2.4 Semantic Annotation for Automatic Generation of Domain-Specific Annotation Sets

We focus on ontology-based annotation techniques which link pre-processed forms of textual content to a specific concept or relation of a domain-specific ontology (which is integrated in the domain-specific layer of *LinkedMDR*). This is to automatically generate domain-specific annotation sets that will be later used to create *LinkedMDR* domain-specific instances.

To do this, we also use the GATE API as it provides robust techniques for automatic semantic annotation and has been frequently used in the literature for this purpose [23, 56]. As it is the case with the use of GATE in the automatic generation of inter and intra-document references, *LMDR Annotator* executes a main pre-configured GATE pipeline application. For this sub-module, the GATE pipeline is made of the main following processing resources:

- *Document Reset*, *Tokeniser*, *Sentence Splitter*, *POS Tagger*, and *Morphological Analyser* for linguistic pre-processing of the textual content;
- *FlexibleGazetteer* where it is possible to configure (i) feature names of annotation sets that will replace the original text, and (ii) a gazetteer instance which generates ontology-based annotations with the configured features. The latter comes down to an *OntoRootGazetteer* which is fed with the chosen domain-specific ontology in order to pre-process its classes, properties, labels, etc. (using the same above linguistic processing resources) and generate their human-understandable forms.

LMDR Annotator outputs the generated annotation sets in XML format. An extract of the generated output following the execution of the main GATE pipeline, described above, is illustrated in Fig. 4.8. For instance, the concept *SolarFactor* is identified from the occurrence of “facteur solaire" in the French textual context of a construction related document.

```
<Annotation Id="37" Type="Lookup" StartNode="0" EndNode="15">
<Feature>
  <Name className="java.lang.String">heuristic_value</Name>
  <Value className="java.lang.String">facteur solaire</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">majorType</Name>
  <Value className="java.lang.String"></Value>
</Feature>
<Feature>
  <Name className="java.lang.String">type</Name>
  <Value className="java.lang.String">class</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">propertyURI</Name>
  <Value className="java.lang.String">http://www.w3.org/2000/01/rdf-schema#label</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">URI</Name>
  <Value className="java.lang.String">http://spider.sigappfr.org/linkedmdr/#SolarFactor</Value>
</Feature>
```

FIGURE 4.8 – Extract of the generated domain-specific annotation sets using GATE for automatic semantic annotation.

4.2.5 *LinkedMDR* Converters

LinkedMDR Converter is the core sub-module of *LMDR Annotator* as it is made of a list of converters that bridge the gap between annotations of existing standards and tools, and those of *LinkedMDR* over heterogeneous documents. The main purpose of this sub-module is to automatically generate instances of *LinkedMDR* based on the outputs of all previously described sub-modules.

Each sub-module in *LMDR Annotator* has a dedicated *LinkedMDR* converter (See Fig. 4.1). The latter is responsible for the conversion of an XML file into an RDF file describing instances of *LinkedMDR* ontology. The advantage of *LinkedMDR* converters is that, even if we modify the underlying APIs and web services of adjacent sub-modules generating the XML files, the converters remain the same.

The conversions do not consist of simple transformations of XML descriptors into RDF instances. They further create the required *LinkedMDR* core instances that are not generated by any of the existing APIs or web services. These core instances ensure the connection among other instances generated based on the existing descriptors.

In order to do the required conversions, *LMDR Annotator* relies on eXtensible Stylesheet Language Transformations[14] (XSLT) as it is the widely used language to transform XML documents into other formats. Consequently, each converter comes down to a tailored XSLT processor:

- *DC to LinkedMDR* is an XSLT processor that transforms DC descriptors into *LinkedMDR* instances based on the semantics provided by the Core Layer (See Sect. 2.3.2.1) and the Standardized Metadata Layer, more precisely the sub-layer dedicated to DC standard (See Sect. 2.3.2.2);

[14] A W3C Recommendation, `https://www.w3schools.com/xml/xsl_intro.asp`

- *TEI to LinkedMDR* is an XSLT processor that transforms TEI descriptors into *LinkedMDR* instances based on the semantics provided by the Core Layer and the Standardized Metadata Layer, more precisely the sub-layer dedicated to TEI standard;
- *MPEG-7 to LinkedMDR* is an XSLT processor that transforms DC descriptors into *LinkedMDR* instances based on the semantics provided by the Core Layer and the Standardized Metadata Layer, more precisely the sub-layer dedicated to MPEG-7 standard;
- *xLinks to LinkedMDR* is an XSLT processor that transforms GATE annotation sets describing inter and inter-document references into *LinkedMDR* instances based on the semantics provided by the Core Layer;
- *Domain to LinkedMDR* is an XSLT processor that transforms GATE annotation sets describing domain-specific information into *LinkedMDR* instances based on the semantics provided by the Core Layer and the pluggable Domain-Specific Layer (See Sect. 2.3.2.3).

An example of a *LinkedMDR* converter, *DC to LinkedMDR*, is provided in Appendix A together with the RDF output generated when executing it on an XML file. The chosen XML file is the same one presented in Fig. 4.3b, which is considered a previously generated output from the Automatic Metadata Extraction sub-module of *LMDR Annotator*. Note that, all the DC descriptors that are contained in the XML file were transformed into *LinkedMDR* instances of the Standardized Metadata Layer (e.g., instance of the concept *dc:title*) in the generated RDF file. Also note that, the converter involves a set of rules that infer *LinkedMDR* instances of the Core Layer (e.g., instance of the concept *Document*) that did not exist in the XML file.

A full list of *LinkedMDR* converters is available online at `spider.sigappfr.org/linkedmdr/lmdr-annotator/`.

4.2.6 Tightly Coupled Semantic Graph Builder

This sub-module is responsible for the aggregation of all the previously generated RDF annotations of *LinkedMDR* converters' sub-module and the generation of a tightly coupled semantic graph $\mathcal{G}_{\delta_m}$ representing a heterogeneous document corpus δ_m.

We use the Apache Jena 3.2.0 API[15] as it is the widely used API to extract data from and write to RDF graphs. On the basis of Jena features, *LMDR Annotator* is capable of:

- Merging RDF files into one file describing the graph $\mathcal{G}_{\delta_m}$, while easily handling RDF instances contained in different RDF files and referring to the same resource, based on the latter's Universal Resource Identifier (URI);

[15] Available at `https://jena.apache.org/`

- Creating hybrid edges which link *LinkedMDR* domain-specific instances to core instances or standardized metadata-based instances where they were identified during the semantic annotation process;
- Dynamically generating inferred instances in $\mathcal{G}_{\delta_m}$ based on the semantics of *LinkedMDR* ontology.

FIGURE 4.9 – Extract of RDF file describing $\mathcal{G}_{\delta_m}$ related to a heterogeneous document corpus δ_m involving 8 construction related documents.

FIGURE 4.10 – Extract of RDF file illustrated in Fig. 4.9 with a zoom in describing *LinkedMDR* document instance.

Fig. 4.9 shows an extract of an RDF file describing a given heterogeneous corpus δ_m containing 8 different construction related documents. Fig. 4.10 shows an extract of a zoom in describing a PDF construction related document. One can see that, for the same document, there are *LinkedMDR* instances generated based on descriptors of the different previous *LMDR Annotator* sub-modules (e.g., Automatic Metadata Generation of DC and TEI Metadata). For instance, the *LinkedMDR* instance describing the PDF document has several document properties: *LinkedMDR* instances describing general metadata of the document (e.g., *dc:title* describing the title of the document), and *LinkedMDR* instances describing text metadata (e.g., *tei:div* describing a section of the same document).

4.3 *HM Query Processor*

We developed *HM Query Processor*, which stands for ***H***ybrid ***M***olecule-based ***Query Processor***, within *FEED2SEARCH* Java-based prototype. Its main purpose is to retrieve a ranked list of relevant hybrid molecule-based query answers w.r.t. a given user's natural language query. Its architecture conforms to the *Hybrid Molecule-based Query Processing* Layer of *FEED2SEARCH* (See Fig 1.5 in Chapter 1). *HM Query Processor* is the implementation of *HM Query Processing* algorithms detailed in Chapter 3:

- *HM_QueryInterpretation* for hybrid molecule-based query interpretation module (Sect. 3.4.2.1), which extracts domain-specific concept occurrences from the user's natural language query;
- *HM_Search* for hybrid molecule-based search module (Sect. 3.4.2.2), which traverses the RDF graph starting from nodes matching the previously extracted domain-specific occurrences and searches for relevant hybrid molecules;
- *HM_Ranking* for hybrid molecule-based ranking module (Sect. 3.4.2.3), which assigns scores to the extracted hybrid molecules and ranks them in descending order;
- *HM_Presentation* for hybrid molecule-based presentation module (Sect. 3.4.2.4), which displays the ranked list of hybrid molecules and shows, for each hybrid molecule query answer, its core information, its domain-specific and structural-based context.

For the implementation of these algorithms, *HM Query Processor* uses:

- Two of *LMDR Annotator*'s sub-modules for the query interpretation: the semantic annotation sub-module relying on Gate Developer 8.1 API (See Sect. 4.2.4) and *LinkedMDR* Converters sub-module (See Sect. 4.2.5) in order to use the same semantic annotation technique as for the corpus annotation phase and then transform the annotations into *LinkedMDR* instances;
- Jena 3.2.0 API, which provides a means to navigate within the semantic graph, extract data and reason over it in the search and ranking phases;
- NavigOWL [16] java-based visualization tool which is called by *HM Query Processor* in the presentation phase, specifically for the visualization of the domain-specific and structural-based context of each hybrid-molecule query answer in an appealing graph layout [52]. The tool supports both RDF and OWL files.

Note that, the implementation of *HM Query Processor* is still on-going, particularly for the presentation module. Our current effort focuses on the GUI in order to improve the presentation of the hybrid molecules and make it more adapted to non computer expert users.

[16] Available at `http://home.deib.polimi.it/hussain/navigowl/`.

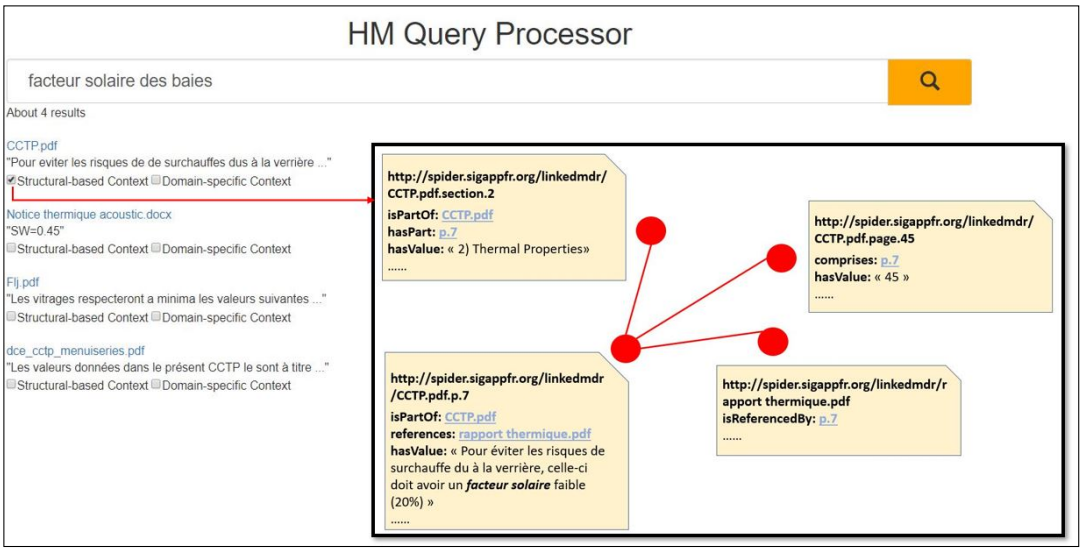


FIGURE 4.11 – Example of the desired *HM Query Processor*'s GUI.

Fig. 4.11 shows an example of the GUI that we are currently working on. The query results are hybrid molecules retrieved following the user's natural language query, in French plain text, regarding the solar factors of windows. For each answer, the user first gets information from the core of the hybrid molecule: the relevant document (e.g., *CCTP.pdf*) and the value of the core's structural-based node of the molecule containing information regarding the solar factor (e.g., *"Pour éviter les risques de surchauffes dus à la verrière, celle-ci doit avoir un facteur solaire faible (20%)"*). The user has the choice to explore the structural-based and the domain-specific context of this answer by checking one or both checkboxes below the answer. This displays the required contextual information, in a graph layout, in the NavigOWL tool. The user can manage the graph in a more clear and consistent view. When the mouse is over a particular node, he can see its value and relations to other nodes within the context (e.g, a reference from the relevant paragraph of *CCTP.pdf* describing the solar factor to another document (*rapport thermique.pdf*) that was not listed in the result list.

4.4 Evaluation of the Annotation of a Heterogeneous Document Corpus based on *LinkedMDR*

We conducted several experiments in order to assess the annotation of a heterogeneous document corpus based on the proposed infrastructure of *LinkedMDR* ontology (See Chapter 2). This is done regardless of the pluggable Domain-Specific Layer as we focus on the invariant knowledge part of *LinkedMDR* across domains. To do so, we targeted two main objectives:

- *Objective 1.1:* Compare *LinkedMDR*'s comprehensive annotations with alternatives ones (i.e., annotations based on existing standards e.g., DC [29], TEI [98],

MPEG-7 [97], and the naive combination of their annotations[17]) regardless of the annotation tools;

- *Objective 1.2:* Evaluate the automatically generated *LinkedMDR* annotations using *LMDR Annotator* with a focus on those generated following the Core Layer and the Standardized Metadata Layer.

4.4.1 Experimental Context

In this section, we present the test corpora used throughout this experimental study together with their corresponding annotations considering different scenarios for *Objective 1.1* and *Objective 1.2*.

4.4.1.1 Test Corpora

We hand-picked 5 heterogeneous document corpora (δ_m with $m \in \{1, \ldots, 5\}$) of real-world construction projects from Nobatek/INEF4. Table 4.1 shows the document composition of each test corpus. Each corpus is made of different documents, where we varied their number, their formats and their size.

TABLE 4.1 – Document composition of the 5 test corpora.

Corpus (δ_m)	No. of Documents	Document Formats	Corpus Size (MB)
$m = 1$	10	3 docx, 4 pdf, 3 png	20
$m = 2$	10	7 pdf, 2 png, 1 jpeg	27.4
$m = 3$	17	5 docx, 10 pdf, 2 png	54.8
$m = 4$	15	5 xslx, 1 docx, 7 pdf, 2 png	112.3
$m = 5$	12	1 xslx, 10 pdf, 1 png	38.2

4.4.1.2 Test Annotations

We prepared two categories of annotations (i) comprehensive annotations representing corpus δ_1 (for *Objective 1.1*), and (ii) automatically generated *LinkedMDR* annotations for the remainder corpora i.e., corpora δ_2, δ_3, δ_4 and δ_5 (for *Objective 1.2*).

As for the comprehensive annotations, we first used annotations generated by *LMDR Annotator* sub-modules on corpus δ_1 (i.e., $\delta_1[DC].xml$, $\delta_1[TEI].xml$, $\delta_1[MPEG7].xml$, and $\mathcal{G}_{\delta_1}.rdf$), and then manually adapted them. This ensured the best possible representation of δ_1 following DC [29], TEI [98], MPEG-7 [97] and *LinkedMDR* (excluding Domain-Specific Layer's annotations). We further combined the generated XML annotations of the three standards in one XML file (i.e., $\delta_1[combined].xml$) to create comprehensive annotations of δ_1 considering their naive combination.

As for the automatically generated *LinkedMDR* annotations, we ran *LMDR Annotator* on the four document corpora: δ_2, δ_3, δ_4 and δ_5. For each document corpus δ_m,

[17] A simple aggregation of the annotations.

we collected the generated RDF Annotations of the different sub-modules separately (i.e., $\delta_m[LinkedMDR\_xLinks].rdf$) representing *LinkedMDR* instances related to inter and intra-document references, and $\delta_m[LinkedMDR\_DC].rdf$, $\delta_m[LinkedMDR\_TEI].rdf$ and $\delta_m[LinkedMDR\_MPEG7].rdf$ representing *LinkedMDR* instances related to the standardized metadata layer. We further collect the generated tightly coupled semantic graph (i.e., $\mathcal{G}_{\delta_m}.rdf$) representing all the generated *LinkedMDR* instances.

4.4.2 Evaluation Criteria and Metrics

In this section, we present the evaluation criteria and the underlying metrics. For *Objective 1.1*, we choose the conciseness criterion that we explain in Sect. 4.4.2.1 which is a well-known criterion among the ontology evaluation methods [77]. We adapt it to our context and define the required metrics in order to take into consideration the evaluation of different data models (ontology-based and non ontology-based) which have different formats and semantics. For *Objective 1.2*, we choose the effectiveness that we explain in Sect. 4.4.2.2 as it is commonly used to evaluate the quality of systems and tools' results in the context of IR [63].

4.4.2.1 Conciseness

We evaluate the conciseness of each adopted annotation model i.e., DC [29], TEI [98], MPEG-7 [97], the three standards combined, and *LinkedMDR* ontology in the representation of corpus δ_1. More particularly, for each annotation model we look at:

- The number of annotated documents;
- The number of resulting annotation files;
- The cumulative number of annotation elements (i.e., the number of XML tags in the XML annotation files and the number of RDF triples in the RDF annotation file) within the annotation files;
- The number of redundancies i.e., the overlapping annotation elements;
- The percentage of covering a pre-defined list of relevant criteria. This list comes down to categories of annotation elements aligned with the requirements we set in Chapter 2 such as inter and intra-document links, general metadata of the documents, structural metadata for the text, structural metadata for the image, etc. (See Sect. 2.1).

We consider an annotation model to be the most concise if it is capable of annotating all the documents in the corpus δ_1 with a minimum number of annotation elements, a minimum number of annotation files and a minimum number of redundancies, while covering a maximum number of relevant criteria from the pre-defined list.

4.4.2.2 Effectiveness

We evaluate the effectiveness of *LMDR Annotator* in automatically annotating corpora δ_2, δ_3, δ_4, and δ_5 in terms of:

- *Precision* (P) to identify the number of relevant annotations among the automatically generated ones;
- *Recall* (R) to identify the number of relevant automatically generated annotations among the total number of expected relevant annotations;
- F_2-score to evaluate a weighted average of P and R. Note that we choose F_2-score since it emphasizes the recall measure. This indirectly highlights missed expected relevant annotations, which is an important factor to consider in the evaluation of the quality of the annotation phase.

These measures are calculated as follows:

$$P = \frac{a}{a+b} \tag{4.1}$$

$$R = \frac{a}{a+c} \tag{4.2}$$

$$F_2 - Score = \frac{5 \times P \times R}{4 \times P + R} \tag{4.3}$$

Where,

- a is the number of automatically generated relevant annotations (true positives);
- b is the number of automatically generated annotations that are not relevant (false positives).
- c is the number of relevant annotations that are not generated by *LMDR Annotator* (false negatives).

4.4.3 Experimental Results

In this section, we present the results of two conducted experiments considering the context defined in Sect. 4.4.1, and the evaluation criteria and metrics defined in Sect. 4.4.2.

4.4.3.1 Evaluating the conciseness of *LinkedMDR* and its alternatives

We evaluate the conciseness of *LinkedMDR*, DC [29], TEI [98], MPEG-7 [97], and the three standards combined in annotating δ_1 as described in Sect. 4.4.2.1.

Table 4.2 shows that when using DC, the three standards combined, and *LinkedMDR* as annotation models, all the documents involved in corpus δ_1 were annotated. In contrast, when using TEI and MPEG-7 as annotation models, the annotations were not exhaustive. This is due to the incapacity of annotating images and technical drawings in TEI and textual documents in MPEG-7. The annotation files generated based on DC standard contain a very small number of annotation elements while covering few relevant criteria from the pre-defined list (21%). This is because DC covers generic metadata and neglects structure and content representation of the documents. The naive combination of the three standards produces a significant number of annotation elements since TEI and MPEG-7 standards are very verbose while covering 77% of relevant pre-defined criteria. Examples on the important relevant criteria that the annotations of the three standards combined still lack are inter and intra-document dependencies. Also, the naive combination of the three standards involves many redundancies caused by mutual annotation elements between DC, TEI and MPEG-7 standards, mainly general metadata information of the documents (e.g., document title, author name, etc.).

The current experiment shows that *LinkedMDR* is the most concise annotation model representing corpus δ_1 since it has the highest coverage of relevant pre-defined criteria (100%) with a relatively small number of annotation elements, all in a single annotation file, and without any redundancies.

TABLE 4.2 – Evaluating the conciseness of the existing standards and *LinkedMDR* in annotating corpus δ_1.

Annotation Model	No. of Annotated Documents	No. of Annotation Files	Cumulative No. of Annotation Elements	No. of Redundancies	Coverage of Relevant Criteria
DC [29]	10	10	131	0	21%
TEI [98]	5	5	807	0	49 %
MPEG-7 [97]	5	5	495	0	28%
Baseline	10	20	1433	128	77%
LinkedMDR	10	1	604	0	100%

4.4.3.2 Evaluating the Effectiveness of *LMDR Annotator*

We evaluate the effectiveness of *LMDR Annotator* in automatically annotating corpora δ_2, δ_3, δ_4, and δ_5 as described in Sect. 4.4.2.2.

Fig. 4.12 shows F_2-scores evaluating the outputs of each *LMDR* sub-module separately (i.e., $\delta_m[LinkedMDR\_DC].rdf$, $\delta_m[LinkedMDR\_TEI].rdf$, $\delta_m[LinkedMDR\_MPEG7].rdf$, and $\delta_m[LinkedMDR\_xLinks].rdf$) then their union (i.e., $\mathcal{G}_{\delta_m}.rdf$ excluding the domain-specific part i.e., $\delta_m[LinkedMDR\_Domain].rdf$). In general, F_2-scores for the overall *LMDR Annotator*'s sub-modules range from 0.48 (for $\mathcal{G}_{\delta_4}.rdf$) to 0.63 (for $\mathcal{G}_{\delta_2}.rdf$).

Looking over the individual annotation sub-modules, F_2-scores slightly change from one corpus to another. *LMDR Annotator*'s sub-module generating $\delta_m[LinkedMDR\_DC].rdf$ is, in general, the most effective one since it involves *LinkedMDR* instances generated from documents' meta-tags (e.g., title, creator, date, format, etc.) which are easy to extract automatically using the Apache Tika API (See Sect. 4.2.2.1). On the other

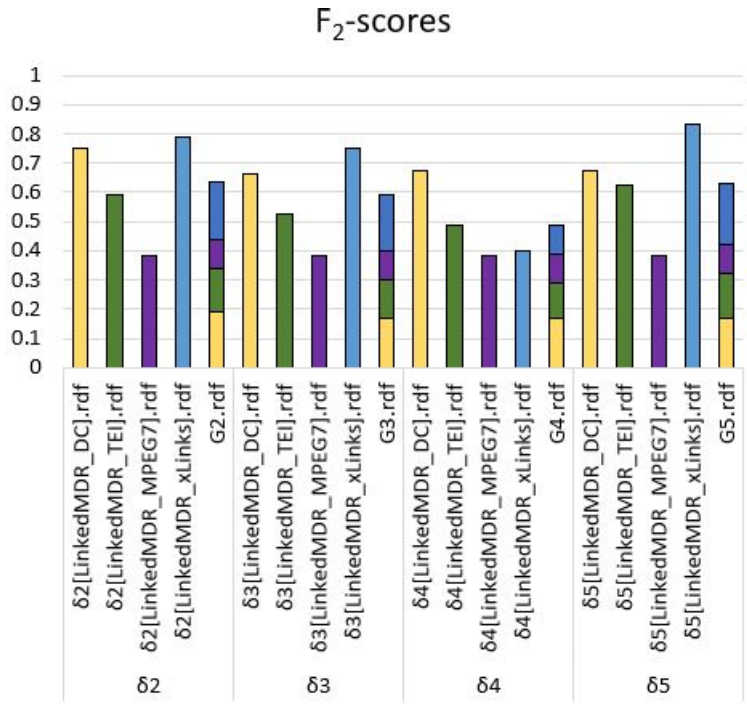


FIGURE 4.12 – F_2-scores measuring the effectiveness of LMDR Annotator in annotating corpora δ_2, δ_3, δ_4, and δ_5.

hand, the sub-module generating $\delta_m[LinkedMDR\_MPEG7].rdf$ produces the lowest scores since relevant concepts, such as *mpeg7:StillRegion*, are relatively difficult to generate automatically. In fact, *LMDR Annotator* uses the MPEG-7 Visual Descriptors library, which is limited to the automatic extraction of low level features, such as color and texture characteristics (See Sect. 4.2.2.3). Advanced feature extraction requires sophisticated computer vision and machine learning algorithms (such as [71]) which, so far, are not adopted in our prototype. The sub-module generating $\delta_m[LinkedMDR\_TEI].rdf$ provides relatively good results in automatically generating structural metadata related to the text using Apache Oxgarage web service (See Sect. 4.2.2.2). However, there are still some limitations due to originally poor structured textual documents (especially with some PDF files) resulting in missing annotations regarding headings numbering and information on their related pages. As for the sub-module generating $\delta_m[LinkedMDR\_xLinks].rdf$, one can see that *LMDR Annotator* is reliable in automatically extracting explicit inter and intra-document references from pre-defined regular expressions encountered in the textual documents using the Java GATE API (See Sect. 4.2.3). However, in some cases, such as in corpus δ_4, the sub-module obtains relatively lower score. This is due to some expressions representing ambiguous references that could not be handled automatically without the use of advanced semantic disambiguation techniques [21, 94].

Fig. 4.13 illustrates the recall (R) values w.r.t. to the total expected LinkedMDR instances per corpus. This is to compare current state of *LMDR Annotator* in automatically generating *LinkedMDR* instances with the annotation potential proposed

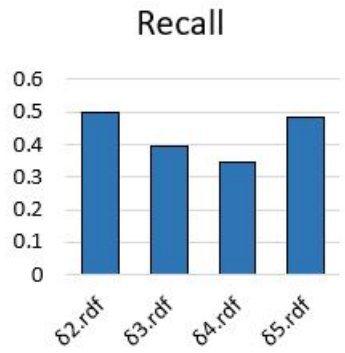


FIGURE 4.13 – Recall (R) scores based on the total expected *Linked-MDR* instances per corpus δ_m.

by *LinkedMDR* ontology. The average recall value is equal to 0.43 over the four corpora. Intuitively, the recall scores decrease when the number of documents increases (e.g., from $R = 0.5$ for corpus δ_2 to $R = 0.34$ for corpus δ_4) since more complex inter and intra-document links are involved, which cannot yet be generated by *LMDR Annotator* (e.g., spatial inter-document dependencies: overlaps of images representing extract of technical drawings across documents). This emphasizes that *LMDR Annotator*, at its current state, offers relatively low recall scores when compared to the annotation potential proposed by *LinkedMDR* ontology.

4.4.4 Discussion

The first experiment demonstrates that, considering optimal annotations, the annotation capability provided by *LinkedMDR* ontology is better than the one provided by commonly used existing metadata standards for document representation (*Objective 1.1*).

In a real-world application, manually annotating a document corpus is a tedious job that requires much effort and technical knowledge. The second experiment shows promising results for an automatic annotation of a heterogeneous document corpus using *LMDR Annotator*, which automatically generates *LinkedMDR* instances representing the corpus (*Objective 1.2*). Although the average obtained effectiveness value is still low (average F_2-score$= 0.6$), it reflects the quality of the so far implemented sub-modules of *LMDR Annotator* mostly relying on existing APIs, tools and services embedded in an automatic pipeline. However, these results are still encouraging since the pipeline provides more automatic annotation capabilities over a given corpus than each existing API, tool or web service alone. Our current effort focuses on furthering the annotation capabilities of *LMDR Annotator* in order to obtain higher effectiveness results enabling its adoption in real-world projects.

4.5 Evaluation of the Quality of the Proposed Hybrid Molecule-based Search and Ranking

We conducted several experiments in order to assess the quality of the retrieved hybrid molecule-based query answers using the proposed *HM_CSA* algorithm and weight mapping functions. This is done using *HM Query Processor* described in Sect 4.3. To do so, we targeted two main objectives:

- *Objective 2.1:* Validate that *HM_CSA* can provide relevant hybrid molecules w.r.t. users' queries;
- *Objective 2.2:* Validate that the *HM_Ranking* module can rank the generated hybrid molecules conveniently w.r.t. users' queries.

4.5.1 Experimental Context

In this section, we describe the users' queries together with the test data used throughout this experimental study for *Objectives 2.1* and *2.2*.

4.5.1.1 Queries

We collected 25 queries from Nobatek/INEF4 based on frequently required information searched by actors with different expertise (architects, technicians and engineers) throughout the different stages of real-world construction projects. We divided the queries into two groups:

- Query Group 1: $q_1 \rightarrow q_{12}$ for simple queries i.e., queries firing start nodes of 1 concept type. For instance, q_1= *"Charateristiques des chassis vitrés"* fires 1 domain-specific concept type which is *IfcCurtainWall*;
- Query Group 2: $q_{13} \rightarrow q_{25}$ for more diverse queries i.e., queries firing start nodes of 2 or more concept types. For instance, q_{13}= *"Facteur solaire des baies"* fires 2 domain-specific concept types which are *SolarFactor* and *IfcWindow*.

4.5.1.2 Test Data

We generated a tightly coupled semantic graph δ_6 of 30 000 RDF triples over a heterogeneous document corpus δ_6 of 15 documents hand-picked from a real-world construction project in Nobatek/INEF4 (See Table 4.3). The graph was first generated using the implemented part of *LMDR Annotator* (See Sect. 4.2).

In order to asses the quality of the search and ranking modules regardless of the annotation tool that generates the graph, we manually completed the annotations in order to ensure the best representation of the document corpus.

TABLE 4.3 – Heterogeneous document corpus δ_6.

No. of Documents	Format	Content Description
5	DOCX	General Technical Specifications
		Electrical Specifications
		Exterior Facades and Carpentry
		Thermal Properties
		Acoustic Properties
7	PDF	Electrical Drawing
		HVAC Drawing
		Wall Composition
		Confort Analysis
		Environmental Impacts
		Environmental and Energy Impacts
		Carpentry and Glazing
1	XLSX	Thermal Regulations
2	PNG	Material Pattern Photo
		Sealing Test

4.5.2 Evaluation Criterion and Metrics

We choose the effectiveness of the retrieval as evaluation criterion since it is the most renowned in the evaluation of the retrieval results of an IR system [63]. In this section, we present the different metrics we use in order to evaluate the effectiveness in two different scenarios: (i) regardless of the order of the hybrid molecule-based query answers (for *Objective 2.1*), and (ii) considering their order (for *Objective 2.2*).

As for the first scenario, we evaluate the effectiveness of *HM_CSA* algorithm in terms of:

- *Precision* (P) to identify the number of relevant hybrid molecules among the retrieved results;
- *Recall* (R) to identify the number of relevant hybrid molecules that are retrieved among the total number of expected relevant results;
- F_1-score to evaluate the harmonic mean of P and R, which is the most used in the evaluation of IR results [63].

P and R are calculated following Equations 4.1 and 4.2 respectively (See Sect 4.4.2.2), with a being the number of correctly retrieved hybrid molecules (true positives), b the number of wrongly retrieved hybrid molecules (false positives), and c the number of hybrid molecules that are not retrieved although they are relevant (false negatives). F_1-score is calculated as follows:

$$F_1 - Score = \frac{2 \times P \times R}{P + R} \tag{4.4}$$

As for the second scenario, we evaluate the effectiveness of *HM_Ranking* algorithm on the basis of the proposed weight mapping functions in terms of the Mean Average Precision (*MAP*) measure, which assesses the ranking of relevant hybrid molecule-based query answers. It is calculated as follows:

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \tag{4.5}$$

Where,

- q is a user query;
- Q is the total number of users' queries;
- $AveP(q)$ is the average precision of query q, such that:

$$AveP(q) = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{a + c} \tag{4.6}$$

with $P(k)$ being the precision at rank k, $rel(k)$ equal to 1 if the kth hybrid molecule is relevant and 0 otherwise.

Note that, for the assessment of the relevance of a hybrid molecule used in the calculation of the above metrics, we relied on users' judgments. For each query, we asked 12 users (who did not take part in the queries formulation), highly involved in the construction project from which the documents of the corpus δ_6 were taken, to provide a score for each answer (1 for relevant and 0 for not relevant) independently from each other. Afterwards, the users were asked to validate the judgments collectively. These users also provided the false negative hybrid molecules[18] for each query.

4.5.3 Experimental Scenarios and Results

In this section, we present the results of two conducted experiments considering the context defined in Sect. 4.5.1, and the evaluation criterion and metrics defined in Sect. 4.5.2.

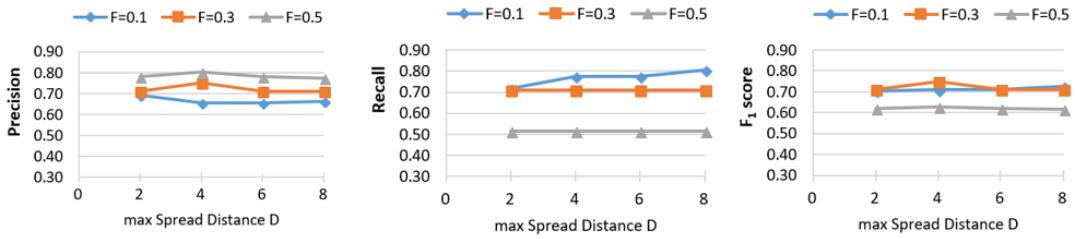
4.5.3.1 Evaluating the effectiveness of *HM_CSA*

We evaluate the effectiveness of *HM_CSA* for the 25 given queries in terms of P, R and F_1-score (*Objective 2.1*). We study the impact of constraint parameters, mainly the firing threshold F and the maximum spread distance D, on the query answers. To do so, we consider 3 different values for F (0.1, 0.3, and 0.5) and 4 different values for D (2,4,6, and 8). This resulted in 12 run configurations for each query execution. We further examine the influence of the diversity of the queries considering

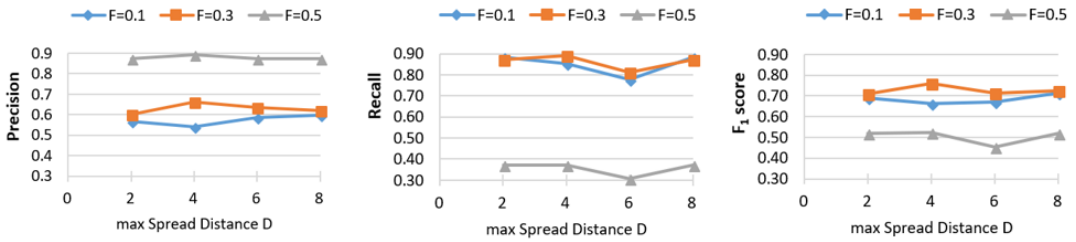
[18] For the sake of simplicity, they only pointed missing hybrid molecules' cores.

the two groups. Figure 4.14 shows the average values of P, R, and F_1-score per run configuration per query group.

We select the optimal values of constraint parameters based on the results of the F_1-score. Figure 4.14 shows that the highest average values of F_1-score for both Query Group 1 (See Figure 4.14a) and Query Group 2 (See Figure 4.14b) are attained with $F = 0.3$ and $D = 4$ (F_1-score$\simeq$ 0.75). The optimal values of the constraint parameters portray a trade off between P and R. For instance, high precision is achieved with higher F values as only the most relevant nodes (with very high activation values) are selected. However, a high F value restricts the spread of the activation in the graph resulting in lower recall values. We also notice that, with the optimal constraint parameters ($F = 0.3$ and $D = 4$), *HM_CSA* attains slightly higher average precision and lower average recall with Query Group 1 when compared to Query Group 2. This is because increasing the concept types of the start nodes ensures that larger portion of the graph is searched but at the cost of increased false positive results.



(A) Query Group 1.



(B) Query Group 2.

FIGURE 4.14 – Average *Precision (P)*, *Recall (R)*, and F_1-*score* of *HM_CSA* considering different values of threshold F and maximum spread distance D for (a) Query Group 1 and (b) Query Group 2.

4.5.3.2 Evaluating the effectiveness of *HM_Ranking*

We also evaluate the ranking of the hybrid molecule-based query answers considering the optimal constraint parameters of *HM_CSA* ($F = 0.3$ and $D = 4$ from the previous experiment) over the same two groups of queries. We vary α and β parameters (See Section 3.4.4.3) and study their impact on the *MAP* values of *HM_Ranking*. We choose 3 configurations: (i) $\alpha = 0$ and $\beta = 1$, (ii) $\alpha = 1$ and $\beta = 0$, and (iii) $\alpha = 1$ and $\beta = 1$ to emphasize respectively domain-specific nodes' contribution,

structural-based nodes' contribution and the equal contribution of both in the overall weight of a hybrid molecule. These contributions also reflect the impact of the different strategies adopted for the weight mapping (See Section 3.4.4). Figure 4.15 shows average values of *MAP* per configuration per group.

The results show that considering only the contribution of domain-specific nodes (i.e., $\alpha = 0$, $\beta = 1$), Query Group 2 attains a higher average *MAP* value in comparison to the result obtained when considering only the contribution of structural-based nodes (i.e., $\alpha = 1$, $\beta = 0$). Query Group 1 shows an opposite behavior as it has less concept types for starting nodes, thus it is less influenced by the domain-specific contribution. The highest values of *MAP* (*MAP*= 0.62 for Query Group 1, *MAP*= 0.74 for Query Group 2) are reached when taking into account both the structural-based and domain-specific aspects in the weight calculation of the molecules (i.e., $\alpha = 1$, $\beta = 1$). This further highlights the importance of the hybrid aspect of the molecule.

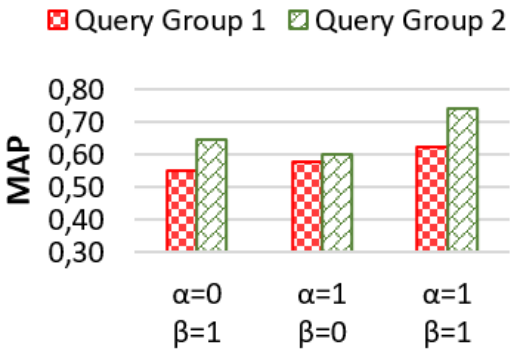


FIGURE 4.15 – Average *MAP* values of *HM_Ranking* per α and β configuration per Query Group.

4.5.4 Discussion

In the context of the given heterogeneous document corpus, the two conducted experiments demonstrate that, using optimal constraint and weight parameters, *HM_CSA* reaches an overall F_1-score of 0.75, and *HM_Ranking* reaches overall *MAP* values > 0.6. The obtained results are considered as promising results in IR [63]. They demonstrate that our proposed Hybrid Molecule-based Search and Ranking modules are capable of providing relevant query answers in the form of hybrid molecules i.e., the augmented contextualized results satisfy the user's needs.

4.6 Summary

This chapter presents an experimental evaluation of the contributions of Chapter 2 and Chapter 3 in regard to the semantic representation of a heterogeneous document corpus based on *LinkedMDR* ontology, and the adoption of such representation in an innovative IR model respectively. In this study, we used real-world data from projects in the AEC industry.

On the one hand, we assessed the annotations of a heterogeneous document corpus considering two different scenarios (i) regardless of the annotation tools i.e, based on comprehensive (ideal) annotations, and (ii) using *LMDR Annotator* module of *FEED2SEARCH* Java-based prototype that we implemented to automatically generate the annotations. The corresponding experimental results show that:

- Considering optimal annotation capabilities (ideal annotations), the annotations based on the infrastructure provided by *LinkedMDR* ontology are more concise than those provided by commonly used existing metadata standards (e.g., DC [29], TEI [98], MPEG-7 [97], and the three standards combined) for document representation;

- Considering automatically generated annotations using *LMDR Annotator*, we obtained promising effectiveness results (i.e., average F_2-score$= 0.6$) that still can be improved. However, *LMDR Annotator* provides more automatic annotation capabilities over a given corpus than each existing API, tool or web service alone which are dedicated to a specific document representation based on a specific standard or data model.

On the other hand, we assessed the quality of the retrieved hybrid molecule-based query answers using *HM Query Processor* module of *FEED2SEARCH* Java-based prototype that we implemented to validate that, given a user's natural language query and optimal annotations of a heterogeneous document corpus: (i) the proposed *HM_CSA* algorithm provides relevant hybrid molecules query answers, and (ii) *HM_Ranking* module ranks the generated hybrid molecules conveniently. The corresponding experimental results show that:

- Considering optimal constraint parameters and diverse queries, the effectiveness reached with *HM_CSA* is relatively good (average F_1-score$= 0.75$), which demonstrates that the retrieved hybrid molecules correspond to users expectations. The *Hybrid Molecules* being a novel data structure, we are currently conducting a qualitative study to compare the richness of such data structure w.r.t. the State-of-the-art (e.g., a standard *CSA* algorithm that provides single nodes results).

- Considering a balanced contribution of the structural and domain-specific parts of the hybrid molecules in the calculation of their weights, *HM_Ranking* attains promising results of effectiveness (average *MAP* values > 0.6). Higher effectiveness values can be achieved in the future by investigating further alternative weight mapping functions.

Chapter 5

Conclusion

"The important thing is not to stop questioning. Curiosity has its own reason for existing."
- Albert Einstein

5.1 Recap

In this thesis, we proposed a semantic representation of a heterogeneous document corpus enabling an innovative IR over it. Our proposal enhances the current status of industries in inferring relevant information for the progress of their projects. Our evaluation was based on examples from construction projects in the AEC industry.

In **Chapter 1**, we highlighted the remaining disregarded issue of substituting raw and unstructured information with intelligent data models empowering knowledge reasoning capabilities and enabling the full potential of Industry 4.0. We then gave a particular attention to the Construction 4.0 in the AEC industry which struggled to keep up with the advances of ICT, such as the BIM multi-dimensional model-based process.

We presented a motivating scenario from the context of a multi-disciplinary project (e.g., the construction project), through which the thesis addressed two main challenges: (i) the knowledge representation of a heterogeneous document corpus, and (ii) the innovative information search enabling the users to extract relevant information from raw documents required for their tasks in the project.

The contributions of the thesis were presented within a generic FEED2SEARCH framework and detailed in the remaining chapters.

In **Chapter 2**, we focused on the main challenge of representing the collective knowledge embedded in a heterogeneous document corpus while considering two dimensions: the structural dimension and the domain-specific dimension of the corpus.

We identified the following sub-challenges: (i) representing the various inter and intra-document links within the corpus, (ii) describing general metadata information of the document, its content, and its structural text and multimedia components all

combined, (iii) associating semantics at content and structural levels of the document, (iv) handling document multimodality, and (v) ensuring extensibility. According to the literature that we reviewed, there is no existing standard or data model for document representation that addressed these challenges combined.

We introduced a novel semantic-based approach, entitled *Tight Coupling* approach, which aims to represent the collective knowledge embedded in a heterogeneous document corpus through a *tightly coupled semantic graph*. We then proposed a backbone multi-layered ontology, entitled *LinkedMDR* [20], which provides the required infrastructure to build this graph through core components easily connected to existing standardized metadata standards and easily pluggable to domain-specific ontologies. We then presented a *tight coupling algorithm* which generates the required graph.

In **Chapter 3**, we focused on the main challenge of handling IR over the proposed *tightly coupled semantic graph* while providing the users with innovative search results i.e., augmented search results with meaningful context including both the structural and the domain-specific dimensions of a heterogeneous document corpus.

We identified the following sub-challenges: (i) providing relevant granularity levels of the documents, (ii) providing relevant inter and intra-document dependencies, and (iii) providing a meaningful context for query answers. We reviewed existing IR models including traditional IR models and SIR models and the variety of approaches and systems that implement them. To our knowledge, none of the existing models addressed these challenges combined.

We first proposed a novel data structure for query answers based on well-defined sub-graphs, which we call *Hybrid Molecules*, extracted from a tightly coupled semantic graph. The *Hybrid Molecules* bring in a core information together with helpful contextual information of the documents including both the structural and the domain-specific dimensions. We then provide a comprehensive query processing pipeline, entitled *HM Query Processing*, on the basis of the proposed *hybrid molecules*. Its main purpose is to generate a ranked list of relevant hybrid molecule-based query answers w.r.t. the user's natural language query. We mainly focused on the Search and Ranking modules which consist of: *HM_CSA*, a novel graph-based search algorithm that generates the *hybrid molecules* from the *tightly coupled semantic graph*, and Weight Mapping functions which score the *hybrid molecules* conveniently.

In **Chapter 4**, we focused on the experimental evaluation study that assesses the main contributions of Chapter 2 and Chapter 3 in the context of the AEC industry.

We first introduced the implemented Java-based prototype with its two dedicated modules. The first one, entitled *LMDR Annotator*, automatically annotates a given heterogeneous document corpus and generates the *tightly coupled semantic graph*. The second one, entitled *HM Query Processor*, uses the generated graph to

provide a ranked list of Hybrid Molecule-based answers in response to a user's natural language query.

We then presented a set of experiments conducted on construction projects, provided by Nobatek/INEF4. On the one hand, we compared the representation of a heterogeneous document corpus based on *LinkedMDR* ontology with alternative data models for document representation. We then tested the automatic annotation capabilities of *LMDR Annotator*. On the other hand, we evaluated the search results provided by *HM_SIRP* w.r.t. ground truth provided by users in the AEC industry. The experiments showed promising results that motivate us in further extending the implementation of the prototype and investigating its efficiency for an adoption in real world projects.

5.2 Future Works

There are several future works we can adopt to further improve (i) the contributions of this thesis, and (ii) their validation process. Some of these works are already in progress (mainly works mentioned in Sect 5.2.5, and Sect 5.2.6).

5.2.1 Extending *LinkedMDR*

LinkedMDR ontology was introduced in Chapter 2 as one of the main contributions. Future improvements of *LinkedMDR* include, but are not limited to:

- Integration of descriptors for audio and video (e.g., from MPEG-7 [97], ID3[1], etc.) since the current version (lmdr-v1.owl) mainly focuses on the selection of metadata descriptors for the image and the text;
- Definition of semantic rules (using SWRL[2]) to empower its reasoning capabilities since the current version involves only semantic characteristics for data properties (e.g., Transitive, Symmetric, Inverse properties, etc.) generating inferred relations when the reasoner is started.

5.2.2 Defining Properties and Operations on the *Hybrid Molecules*

The *Hybrid Molecules* were introduced in Chapter 3 as one of the main contributions. A formal definition was provided and exploited by *HM_CSA*, the graph-based search algorithm, in order to construct hybrid molecule-based query answers.

However, the obtained hybrid molecule-based query answers still have some shortcomings such as a huge number of composing nodes, and a great number of overlapping components between them, which can add noises on the presentation of the results). We believe that, defining properties on the hybrid molecules (e.g.,

[1]Standard for MP3 files, `http://id3.org/`
[2]Semantic Web Rule Language, `https://www.w3.org/Submission/SWRL/`

connectivity to other molecules) and operations (e.g., merge overlapping molecules' components and keep the core of the molecule that has the higher connectivity) would improve their quality and provide a better structure for query answers.

5.2.3 Improving the Query Processing

HM Query Processing was introduced in Chapter 3 as a novel pipeline for IR relying on the *Hybrid Molecules*. We focused on *HM_CSA* algorithm and the Weight Mapping functions as the main contributions for the search and ranking modules respectively. Future improvements of *HM Query Processing* include, but are not limited to:

- Adopting advanced disambiguation techniques [94] in the query interpretation module which can leverage initial start nodes for *HM_CSA* algorithm;
- Optimizing the search module by either adopting search parallelism strategies to construct the *Hybrid Molecules* from the graph, or preparing clusters of hybrid molecules before the search is started. The latter can be easily applied since the *Hybrid Molecules* are defined regardless of the context of the search process;
- Including users relevance feedback techniques [47] to improve the search results. In other words, users' feedbacks launch another iteration of *HM_CSA* algorithm with modified initial weights for start nodes.

5.2.4 Improving the Prototypes

LMDR Annotator and *HM_SIRP* were introduced in Chapter 4 for testing the automatic annotation process of a heterogeneous document corpus based on *LinkedMDR* and IR based on the proposed *HM Query Processing* pipeline respectively. Due to development requirements in terms of time and resources, we focused on specific modules when implementing these prototypes. Future improvements of the prototypes include, but are not limited to:

- Automatically generating more *LinkedMDR* instances. For instance, so far, the Automatic Semantic Annotation module of *LMDR Annotator* generates, from textual context, inter and intra-document references. However, some references could be ambiguous and could implicitly refer to documents or parts of documents. We believe that using advanced disambiguation techniques [94]) could help resolving the ambiguity of such references;
- Improving the presentation module for a more user friendly display of the hybrid molecule-based query answers. For instance, the structural-based and domain-specific contextual information of a query answer are still visualized using NavigOwl java-based tool integrated in *HM_SIRP* prototype. We believe

that more display strategies should be adopted in order to improve the presentation of the results and help the non computer expert users to better interpret them. This could be done by exploring state of the art regarding the display of SERP-like results adopted by current search engines and their impact on end users [5].

5.2.5 Extending the Experiments

Chapter 4 focuses on experimental evaluation showing the power of the annotation of a document corpus based on *LinkedMDR*, and the effectiveness of *LMDR Annotator* and *HM_SIRP* prototypes. In order to further validate our contributions in the context of real-world applications, it is crucial to:

- Evaluate the efficiency of *LMDR Annotator* in terms of time and memory resources;
- Evaluate the efficiency of *HM_SIRP* prototype in terms of time and memory resources;
- Compare the richness of the hybrid molecule-based query answers w.r.t. the state-of-art in IR in order to quantify the impact of the contextual information of the results on the user's experience, especially regarding the time needed by the users to interpret and track the results;
- Apply the conducted experiments in the context of other domains (e.g., the medical domain) using another external domain-specific ontology or data model (e.g., MeSH[3]) for *LinkedMDR*'s Pluggable Domain-specific Layer.

5.2.6 On-going Projects

The contributions of this thesis will be integrated and further explored in the following two projects.

5.2.6.1 *BIM4REN* European Project

BIM4REN[4] is a European project which aims at providing a digital ecosystem to facilitate the integration of innovative digital tools, compatible with the BIM model, in the energy renovation process of the buildings. It has just been started in October 2018 and is expected to finish in 2022. It involves 23 partners (SMEs, Universities, etc.), including Nobatek/INEF4, from 10 different countries in Europe.

HM_SIRP prototype would be further enhanced and integrated in the project to help in searching for HVAC data from textual related documents. This is to provide

[3] An example of plugging in the Medical Subject Headings (MeSH) schema in *LinkedMDR* was presented in Chapter 2. Theoretical efforts have been done, yet awaiting real-world projects in the medical domain for experimental tests.

[4] `https://www.ef-l.eu/our-projects/bim4ren/`

empty BIM models (generated from a 3D scan tool during the renovation phase of the building) with the missing information.

5.2.6.2 *E2S* Project

The Energy Environment Solutions (E2S) project[5] aims at providing a generic information system that can be configured, on demand, by any organization in order to integrate multimedia services dedicated to indexing, storing, enriching and handling security of data from various domains, such as the data related to energy and the environment. It is the result of the collaboration between University of Pau & Adour Countries and two national research organizations: National Institute for Agronomy (INRA) and Institute for Research in Computer Science and Automation (Inria). The project has started in 2018 and is expected to finish in 2021.

Contributions related to the semantic representation of a heterogeneous document corpus based on *LinkedMDR* will be integrated in the project and further explored in the context of Big Data and Data Security.

[5] `https://e2s-uppa.eu/en/index.html`

Appendix A

Example of *LinkedMDR* Converter: *DC to LinkedMDR*

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform" xmlns:
    dc="http://purl.org/dc/elements/1.1/" xmlns:lmdr="http://spider.sigappfr.org/
    linkedmdr/#">
  <xsl:output method="xml" version="1.0" encoding="UTF-8" indent="yes"/>
  <xsl:variable name="document" select="substring-after(DCmetadata/dc:identifier/text
      (),'uploads\')" />
  <xsl:variable name="uppercase" select="'ABCDEFGHIJKLMNOPQRSTUVWXYZ'"/>
  <xsl:variable name="lowercase" select="'abcdefghijklmnopqrstuvwxyz'"/>
  <xsl:template match="/">
   <xsl:text disable-output-escaping="yes"><![CDATA[<!DOCTYPE rdf:RDF [
    <!ENTITY lmdr "http://spider.sigappfr.org/linkedmdr/#" >
    <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >]>]]>
  </xsl:text>
    <xsl:text disable-output-escaping="yes">&lt;rdf:RDF xmlns:rdf=&quot;http://www.w3
        .org/1999/02/22-rdf-syntax-ns#&quot;
     xmlns:lmdr=&quot;http://spider.sigappfr.org/linkedmdr/#&quot;
     xmlns:dc=&quot;http://purl.org/dc/terms/#&quot;
     xmlns:rdfs=&quot;http://www.w3.org/2000/01/rdf-schema#&quot;&gt;
      &lt;rdf:Description rdf:about=&quot;&amp;lmdr;</xsl:text><xsl:value-of select="
          $document"/>
      <xsl:text disable-output-escaping="yes">&quot;&gt;
          &lt;rdf:type rdf:resource=&quot;&amp;lmdr;Document&quot;/&gt;</xsl:text>
        <xsl:for-each select="DCmetadata/*">
          <xsl:text disable-output-escaping="yes">
          &lt;lmdr:hasProperty rdf:resource=&quot;&amp;lmdr;</xsl:text><xsl:value-of
              select="concat($document,'.',local-name(.),'.',generate-id(.))"/><xsl:
              text disable-output-escaping="yes">&quot;/&gt;</xsl:text>
        </xsl:for-each>
      <xsl:text disable-output-escaping="yes">
      &lt;/rdf:Description&gt;</xsl:text>
      <xsl:for-each select="DCmetadata/*">
        <xsl:text disable-output-escaping="yes">
        &lt;rdf:Description rdf:about=&quot;&amp;lmdr;</xsl:text><xsl:value-of select
            ="concat($document,'.',local-name(.),'.',generate-id(.))"/>
          <xsl:text disable-output-escaping="yes">&quot;&gt;
```

```
29          &lt;rdf:type rdf:resource=&quot;&amp;dc;</xsl:text><xsl:value-of select="
                concat(translate(substring(local-name(.), 1, 1), $lowercase, $uppercase)
                ,substring(local-name(.),2,string-length(local-name(.))-1))"/><xsl:text
                disable-output-escaping="yes">&quot;/&gt;
30          &lt;lmdr:hasValue&gt;</xsl:text><xsl:value-of select="."></xsl:value-of>
31          <xsl:text disable-output-escaping="yes">&lt;lmdr:hasValue&gt;
32        &lt;/rdf:Description&gt;</xsl:text>
33      </xsl:for-each>
34    <xsl:text disable-output-escaping="yes">&lt;/rdf:RDF&gt;</xsl:text>
35  </xsl:template>
36
37 </xsl:stylesheet>
```

LISTING A.1 – Underlying XSLT Processor of *DC to LinkedMDR* converter

```
1  <?xml version="1.0" encoding="UTF-8"?><!DOCTYPE rdf:RDF [
2     <!ENTITY lmdr "http://spider.sigappfr.org/linkedmdr/#" >
3     <!ENTITY dc "http://purl.org/dc/terms/#" >
4     <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >]>
5   <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
6      xmlns:lmdr="http://spider.sigappfr.org/linkedmdr/#"
7      xmlns:dc="http://purl.org/dc/terms/#"
8      xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
9       <rdf:Description rdf:about="&lmdr;Descriptif APD Tous lots.pdf">
10          <rdf:type rdf:resource="&lmdr;Document"/>
11          <lmdr:hasProperty rdf:resource="&lmdr;Descriptif APD Tous lots.pdf.
                identifier.N65541"/>
12          <lmdr:hasProperty rdf:resource="&lmdr;Descriptif APD Tous lots.pdf.creator.
                N65543"/>
13          <lmdr:hasProperty rdf:resource="&lmdr;Descriptif APD Tous lots.pdf.created.
                N65545"/>
14          <lmdr:hasProperty rdf:resource="&lmdr;Descriptif APD Tous lots.pdf.modified.
                N65547"/>
15          <lmdr:hasProperty rdf:resource="&lmdr;Descriptif APD Tous lots.pdf.format.
                N65549"/>
16          <lmdr:hasProperty rdf:resource="&lmdr;Descriptif APD Tous lots.pdf.title.
                N65551"/>
17      </rdf:Description>
18        <rdf:Description rdf:about="&lmdr;Descriptif APD Tous lots.pdf.identifier.
              N65541">
19          <rdf:type rdf:resource="&dc;identifier"/>
20          <lmdr:hasValue>C:\Users\cnathalie\workspace\lmdr-annotator-v1\uploads\
                Descriptif APD Tous lots.pdf</lmdr:hasValue>
21        </rdf:Description>
22        <rdf:Description rdf:about="&lmdr;Descriptif APD Tous lots.pdf.creator.N65543"
              >
23          <rdf:type rdf:resource="&dc;creator"/>
24          <lmdr:hasValue>Pierre</lmdr:hasValue>
25        </rdf:Description>
26        <rdf:Description rdf:about="&lmdr;Descriptif APD Tous lots.pdf.created.N65545"
              >
27          <rdf:type rdf:resource="&dc;created"/>
```

```
28          <lmdr:hasValue>2015-08-18T06:19:50Z</lmdr:hasValue>
29        </rdf:Description>
30        <rdf:Description rdf:about="&lmdr;Descriptif APD Tous lots.pdf.modified.N65547
              ">
31          <rdf:type rdf:resource="&dc;modified"/>
32          <lmdr:hasValue>2015-08-19T07:17:15Z</lmdr:hasValue>
33        </rdf:Description>
34        <rdf:Description rdf:about="&lmdr;Descriptif APD Tous lots.pdf.format.N65549">
35          <rdf:type rdf:resource="&dc;format"/>
36          <lmdr:hasValue>application/pdf; version=1.4</lmdr:hasValue>
37        </rdf:Description>
38        <rdf:Description rdf:about="&lmdr;Descriptif APD Tous lots.pdf.title.N65551">
39          <rdf:type rdf:resource="&dc;title"/>
40          <lmdr:hasValue>Descriptif APD Tous lots</lmdr:hasValue>
41        </rdf:Description></rdf:RDF>
```

LISTING A.2 – Example of RDF file generated after the execution of *DC to LinkedMDR* converter

Bibliography

[1] Aggarwal, C. C. "Information Retrieval and Search Engines". In: *Machine Learning for Text*. Springer, 2018, pp. 259–304.

[2] Agostinho, C., Dutra, M., Jardim-Gonçalves, R., Ghodous, P., and Steiger-Garção, A. "EXPRESS to OWL morphism: making possible to enrich ISO10303 Modules". In: *Complex Systems Concurrent Engineering*. Springer, 2007, pp. 391–402.

[3] Angles, R. and Gutierrez, C. "Subqueries in SPARQL." In: *AMW* 749 (2011), p. 12.

[4] Arndt, R., Troncy, R., Staab, S., Hardman, L., and Vacura, M. *COMM: designing a well-founded multimedia ontology for the web*. Springer, 2007.

[5] Bajpai, N. and Arora, D. "An Estimation of User Preferences for Search Engine Results and its Usage Patterns". In: *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*. Springer, 2018, pp. 255–264.

[6] Barbau, R., Krima, S., Rachuri, S., Narayanan, A., Fiorentini, X., Foufou, S., and Sriram, R. D. "OntoSTEP: Enriching product model data using ontologies". In: *Computer-Aided Design* 44.6 (2012), pp. 575–590.

[7] Beetz, M., Tenorth, M., and Winkler, J. "Open-ease". In: *ICRA*. IEEE. 2015, pp. 1983–1990.

[8] Belkin, N. J. and Croft, W. B. "Information filtering and information retrieval: Two sides of the same coin?" In: *Communications of the ACM* 35.12 (1992), pp. 29–38.

[9] Belsky, M., Sacks, R, and Brilakis, I. "A framework for semantic enrichment of IFC building models". PhD thesis. Technion-Israel Institute of Technology, Faculty of Civil and Environmental Engineering, 2015.

[10] Benedetti, F., Bergamaschi, S., and Po, L. "Lodex: A tool for visual querying linked open data". In: *CEUR WORKSHOP PROCEEDINGS*. Vol. 1486. ceur-ws. org. 2015.

[11] Berners-Lee, T., Hendler, J., and Lassila, O. "The semantic web". In: *Scientific american* 284.5 (2001), pp. 34–43.

[12] Bill East. *Construction Operations Building Information Exchange (COBIE)*. `http://www.wbdg.org/resources/construction-operations-building-information-exchange-cobie`. 2016, (accessed: 01.08.2018).

[13] Bloechle, J.-L., Rigamonti, M., Hadjar, K., Lalanne, D., and Ingold, R. "XCDF: a canonical and structured document format". In: *International workshop on document analysis systems*. Springer. 2006, pp. 141–152.

[14] Brut, M., Laborie, S., Manzat, A.-M., and Sedes, F. "Integrating Heterogeneous Metadata into a Distributed Multimedia Information System". In: *COGnitive systems with Interactive Sensors* (2009).

[15] Brynjolfsson, E. and McAfee, A. *Race against the machine: How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy*. Brynjolfsson and McAfee, 2012.

[16] buildingSMART. *IFC-Industry Foundation Classes, IFC Releases*. `http://www.buildingsmart-tech.org/specifications/ifc-releases/`. 2016, (accessed: 01.08.2018).

[17] buildingSMART. *Model View Definitions*. `http://www.buildingsmart-tech.org/specifications/ifc-view-definition`. 2017, (accessed: 01.08.2018).

[18] BuildingSMART Linked Data Working Group (LDWG). *ifcOWL Ontology*. `http://openbimstandards.org/standards/ifcowl/`. 2016, (accessed: 01.08.2018).

[19] Buscaldi, D., Bessagnet, M.-N., Royer, A., and Sallaberry, C. "Using the semantics of texts for information retrieval: a concept-and domain relation-based approach". In: *New Trends in Databases and Information Systems*. Springer, 2014, pp. 257–266.

[20] Charbel, N., Sallaberry, C., Laborie, S., Tekli, G., and Chbeir, R. "LinkedMDR: A Collective Knowledge Representation of a Heterogeneous Document Corpus". In: *International Conference on Database and Expert Systems Applications*. Springer. 2017, pp. 362–377.

[21] Charbel, N., Tekli, J., Chbeir, R., and Tekli, G. "Resolving XML Semantic Ambiguity." In: *EDBT*. 2015, pp. 277–288.

[22] Chbeir, R., Luo, Y., Tekli, J., Yetongnon, K., Ibañez, C. R., Traina, A. J., Traina, C., and Al Assad, M. "SemIndex: Semantic-aware inverted index". In: *ADBIS*. Springer. 2014, pp. 290–307.

[23] Chechev, M., Gonzàlez, M., Màrquez, L., and España-Bonet, C. "The patents retrieval prototype in the MOLTO project". In: *WWW*. ACM. 2012, pp. 231–234.

[24] Cohen, P. R. and Kjeldsen, R. "Information retrieval by constrained spreading activation in semantic networks". In: *INFORM PROCESS MANAG* 23.4 (1987), pp. 255–268.

[25] Crestani, F. and Lee, P. L. "Searching the web by constrained spreading activation". In: *INFORM PROCESS MANAG* 36.4 (2000), pp. 585–605.

[26] Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. "A framework and graphical development environment for robust NLP tools and applications." In: *ACL*. 2002, pp. 168–175.

[27] Della Valle, E., Ceri, S., Barbieri, D. F., Braga, D., and Campi, A. "A first step towards stream reasoning". In: *Future Internet Symposium*. Springer. 2008, pp. 72–81.

[28] Ding, L., Finin, T., Peng, Y., Da Silva, P. P., and McGuinness, D. L. "Tracking rdf graph provenance using rdf molecules". In: *ISWC (Poster)*. 2005, p. 42.

[29] Dublin Core Metadata Initiative. *DCMI Metadata Terms*. `http://dublincore.org/documents/dcmi-terms`. 2012, (accessed: 01.08.2018).

[30] Endris, K. M., Galkin, M., Lytra, I., Mami, M. N., Vidal, M.-E., and Auer, S. "MULDER: Querying the Linked Data Web by Bridging RDF Molecule Templates". In: *DEXA*. Springer. 2017, pp. 3–18.

[31] EXIF. *Exchangeable Image File Format for digital still cameras*. `http://www.exif.org/Exif2-2.PDF`. 2002, (accessed: 01.08.2018).

[32] Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., and Motta, E. "Semantically enhanced information retrieval: An ontology-based approach". In: *Web semantics: Science, services and agents on the world wide web* 9.4 (2011), pp. 434–452.

[33] Fokou, G., Jean, S., Hadjali, A., and Baron, M. "Cooperative techniques for SPARQL query relaxation in RDF databases". In: *European Semantic Web Conference*. Springer. 2015, pp. 237–252.

[34] Frosini, R., Calì, A., Poulovassilis, A., and Wood, P. T. "Flexible query processing for SPARQL". In: *Semantic Web* 8.4 (2017), pp. 533–563.

[35] Fuhr, N. "Probabilistic models in information retrieval". In: *The computer journal* 35.3 (1992), pp. 243–255.

[36] Galkin, M., Collarana, D., Traverso-Ribón, I., Vidal, M.-E., and Auer, S. "SJoin: A Semantic Join Operator to Integrate Heterogeneous RDF Graphs". In: *DEXA*. Springer. 2017, pp. 206–221.

[37] Garcia, R. and Celma, O. "Semantic integration and retrieval of multimedia metadata". In: *5th International Workshop on Knowledge Markup and Semantic Annotation*. 2005, pp. 69–80.

[38] Giunchiglia, F., Kharkevich, U., and Zaihrayeu, I. "Concept search". In: *ESWC*. Springer. 2009, pp. 429–444.

[39] Gospodnetic, O. and Hatcher, E. *Lucene*. Manning, 2005.

[40] Gouws, S., Rooyen, G., and Engelbrecht, H. A. "Measuring conceptual similarity by spreading activation over Wikipedia's hyperlink structure". In: *ACL*. 2010, pp. 46–54.

[41] Green Building XML. *gbXML Schema*. `http://www.gbxml.org/Schema_Current_GreenBuildingXML_gbXML`. 2015, (accessed: 01.08.2018).

[42] Greenberg, J. "Metadata extraction and harvesting: A comparison of two automatic metadata generation applications". In: *Journal of Internet Cataloging* 6.4 (2004), pp. 59–82.

[43] Griffith, J., O'riordan, C., and Sorensen, H. "A constrained spreading activation approach to collaborative filtering". In: *KES*. Springer. 2006, pp. 766–773.

[44] Guo, K., Liang, Z., Tang, Y., and Chi, T. "SOR: An optimized semantic ontology retrieval algorithm for heterogeneous multimedia big data". In: *Journal of Computational Science* (2017).

[45] Haag, F., Lohmann, S., Siek, S., and Ertl, T. "QueryVOWL: Visual composition of SPARQL queries". In: *International Semantic Web Conference*. Springer. 2015, pp. 62–66.

[46] Haag, F., Lohmann, S., Bold, S., and Ertl, T. "Visual SPARQL querying based on extended filter/flow graphs". In: *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*. ACM. 2014, pp. 305–312.

[47] Haines, D. and Croft, W. B. "Relevance feedback and inference networks". In: *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1993, pp. 2–11.

[48] Heaps, H. S. *Information retrieval, computational and theoretical aspects*. Academic Press, 1978.

[49] Huang, C.-C. and Lin, S.-H. "Sharing knowledge in a supply chain using the semantic web". In: *Expert Systems with Applications* 37.4 (2010), pp. 3145–3161.

[50] Hunter, J. "An overview of the MPEG-7 Description Definition Language (DDL)". In: *IEEE Transactions on circuits and systems for video technology* 11.6 (2001), pp. 765–772.

[51] Huovila, P. "Linking IFCs and BIM to sustainability assessment of buildings". In: *Proceedings of the CIB W*. Vol. 78. 2012, p. 2012.

[52] Hussain, A., Latif, K., Rextin, A. T., Hayat, A., and Alam, M. "Scalable visualization of semantic nets using power-law graphs". In: *Applied Mathematics & Information Sciences* 8.1 (2014), p. 355.

[53] Jacobsson, M., Linderoth, H. C., and Rowlinson, S. "The role of industry: an analytical framework to understand ICT transformation within the AEC industry". In: *Construction Management and Economics* 35.10 (2017), pp. 611–626.

[54] Jiang, S., Zhang, H., and Zhang, J. "Research on BIM-based Construction Domain Text Information Management." In: *JNW* 8.6 (2013), pp. 1455–1464.

[55] Kara, S., Alan, Ö., Sabuncu, O., Akpınar, S., Cicekli, N. K., and Alpaslan, F. N. "An ontology-based retrieval system using semantic indexing". In: *Information Systems* 37.4 (2012), pp. 294–305.

[56] Kiryakov, A., Popov, B., Terziev, I., Manov, D., and Ognyanoff, D. "Semantic annotation, indexing, and retrieval". In: *JWS* 2.1 (2004), pp. 49–79.

[57] Linked Building Data (LBD) W3C Community Group. *Building Product Ontology (PRODUCT)*. `https://github.com/w3c-lbd-cg/product/`. 2018, (accessed: 01.08.2018).

[58] Linked Building Data (LBD) W3C Community Group. *Building Topology Ontology (BOT)*. `https://w3c-lbd-cg.github.io/bot/`. 2018, (accessed: 01.08.2018).

[59] Linked Building Data (LBD) W3C Community Group. *Ontology for Property Management (OPM)*. `https://github.com/w3c-lbd-cg/opm`. 2018, (accessed: 01.08.2018).

[60] Lu, Y. "Industry 4.0: A survey on technologies, applications and open research issues". In: *Journal of Industrial Information Integration* 6 (2017), pp. 1–10.

[61] Lux, M. and Chatzichristofis, S. A. "Lire: lucene image retrieval: an extensible java cbir library". In: *Proceedings of the 16th ACM international conference on Multimedia*. ACM. 2008, pp. 1085–1088.

[62] Mangold, C. "A survey and classification of semantic search approaches". In: *Int. J. Metadata, Semantics and Ontology* 2.1 (2007), pp. 23–34.

[63] Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to information retrieval*. Cambridge University Press, 2008. ISBN: 978-0-521-86571-5.

[64] Maynard, D., Bontcheva, K., and Augenstein, I. "Natural language processing for the semantic web". In: *Synthesis Lectures on the Semantic Web: Theory and Technology* 6.2 (2016), pp. 1–194.

[65] McArthur, J. "A building information management (BIM) framework and supporting case study for existing building operations, maintenance and sustainability". In: *Procedia engineering* 118 (2015), pp. 1104–1111.

[66] Mihalcea, R. and Moldovan, D. "Semantic indexing using WordNet senses". In: *Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 11*. Association for Computational Linguistics. 2000, pp. 35–45.

[67] Navigli, R. "Word sense disambiguation: A survey". In: *ACM computing surveys (CSUR)* 41.2 (2009), p. 10.

[68] Newforma. *Newforma Project Center Twelfth Edition*. `https://www.newforma.com/`. 2018, (accessed: 07.08.2018).

[69] Newman, A., Li, Y.-F., and Hunter, J. "A scale-out RDF molecule store for improved co-identification, querying and inferencing". In: *SSWS*. 2008.

[70] Ontotext. *Ontotext Platform*. `https://ontotext.com/products/ontotext-platform/`. 2014, (accessed: 07.08.2018).

[71] OpenCV. *Open Source Computer Vision Library*. `http://opencv.org`. 2011, (accessed: 28.09.2018).

[72] Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., and Johnson, D. "Terrier information retrieval platform". In: *European Conference on Information Retrieval*. Springer. 2005, pp. 517–519.

[73] Park, J.-r. and Brenza, A. "Evaluation of semi-automatic metadata generation tools: A survey of the current state of the art". In: *Information technology and libraries* 34.3 (2015), pp. 22–42.

[74] Pauwels, P. and Terkaj, W. "EXPRESS to OWL for construction industry: Towards a recommendable and usable ifcOWL ontology". In: *Automation in Construction* 63 (2016), pp. 100–133.

[75] Pawełoszek, I. "Web 3.0 applications in enterprise strategy". In: *Studia Ekonomiczne* 234 (2015), pp. 129–139.

[76] Plan Transition Numérique dans le Bâtiment (PTNB). *Feuille De Route Opérationnelle*. `http://www.batiment-numerique.fr/uploads/PDF/`. 2014, (accessed: 20.10.2018).

[77] Raad, J. and Cruz, C. "A survey on ontology evaluation methods". In: *Proceedings of the International Conference on Knowledge Engineering and Ontology Development, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. 2015.

[78] Ramkumar, A. S. and Poorna, B. "Ontology Based Semantic Search: An Introduction and a Survey of Current Approaches". In: *Intelligent Computing Applications (ICICA), 2014 International Conference on*. IEEE. 2014, pp. 372–376.

[79] Rasmussen, M. H., Pauwels, P., Hviid, C. A., and Karlshoj, J. "Proposing a central aec ontology that allows for domain specific extensions". In: *Joint Conference on Computing in Construction*. Vol. 1. 2017, pp. 237–244.

[80] Resnik, P. "Using information content to evaluate semantic similarity in a taxonomy". In: *arXiv preprint cmp-lg/9511007* (1995).

[81] Rocha, C., Schwabe, D., and Aragao, M. P. "A hybrid approach for searching in the semantic web". In: *WWW*. ACM. 2004, pp. 374–383.

[82] Russell, S. J. and Norvig, P. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited, 2016.

[83] Saathoff, C. and Scherp, A. "Unlocking the semantics of multimedia presentations in the web with the multimedia metadata ontology". In: *Proceedings of the 19th international conference on World wide web*. ACM. 2010, pp. 831–840.

[84] Salembier, P. and Smith, J. R. "MPEG-7 Multimedia Description Schemes". In: *IEEE Transactions on circuits and systems for video technology* 11.6 (2001), pp. 748–759.

[85] Salton, G., Fox, E. A., and Wu, H. "Extended Boolean information retrieval". In: *Communications of the ACM* 26.11 (1983), pp. 1022–1036.

[86] Salton, G. and McGill, M. J. "Introduction to modern information retrieval". In: (1986).

[87] Scherp, A., EIßING, D., and Saathoff, C. "A method for integrating multimedia metadata standards and metadata formats with the multimedia metadata ontology". In: *International Journal of Semantic Computing* 6.01 (2012), pp. 25–49.

[88] Schevers, H. and Drogemuller, R. "Converting the industry foundation classes to the web ontology language". In: *Semantics, Knowledge and Grid, 2005. SKG'05. First International Conference on*. IEEE. 2005, pp. 73–73.

[89] Scientific and Technical Center Building (CSTB). *Kroqi Platform*. `https://www.kroqi.fr/`. 2018, (accessed: 07.08.2018).

[90] Soylu, A., Kharlamov, E., Zheleznyakov, D., Jimenez-Ruiz, E., Giese, M., and Horrocks, I. "Ontology-based visual query formulation: An industry experience". In: *International Symposium on Visual Computing*. Springer. 2015, pp. 842–854.

[91] Staab, S. and Studer, R. *Handbook on ontologies*. Springer Science & Business Media, 2010.

[92] Suárez-Figueroa, M. C., Atemezing, G. A., and Corcho, O. "The landscape of multimedia ontologies in the last decade". In: *Multimedia tools and applications* 62.2 (2013), pp. 377–399.

[93] Sun, S., Gong, J., He, J., and Peng, S. "A spreading activation algorithm of spatial big data retrieval based on the spatial ontology model". In: *Cluster Comput* 18.2 (2015), pp. 563–575.

[94] Tekli, J., Charbel, N., and Chbeir, R. "Building semantic trees from XML documents". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 37 (2016), pp. 1–24.

[95] Tekli, J., Chbeir, R., Traina, A. J., Traina Jr, C., Yetongnon, K., Ibanez, C. R., Al Assad, M., and Kallas, C. "Full-fledged semantic indexing and querying model designed for seamless integration in legacy RDBMS". In: *Data & Knowledge Engineering* (2018).

[96] Terkaj, W., Pedrielli, G., and Sacco, M. "Virtual factory data model". In: *Proceedings of the Workshop on Ontology and Semantic Web for Manufacturing, Graz, Austria*. 2012, pp. 29–43.

[97] The Moving Picture Experts Group. *MPEG7-Multimedia Content Description Interface*. http://mpeg.chiariglione.org/standards/mpeg-7. 2001.

[98] The Text Encoding Initiative Consortium. *TEI-Text Encoding Initiative*. http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf. 1994, (accessed: 01.08.2018).

[99] Tidd, J., Bessant, J., and Pavitt, K. *Managing innovation integrating technological, market and organizational change*. John Wiley and Sons Ltd, 2005.

[100] Turtle, H. and Croft, W. B. "Inference networks for document retrieval". In: *ACM SIGIR Forum*. Vol. 51. 2. ACM. 2017, pp. 124–147.

[101] UK Cabinet Office and Infrastructure and Projects Authority. *Government Construction Strategy: 2016-2020*. https://www.gov.uk/government/publications/government-construction-strategy-2016-2020. 2016, (accessed: 20.10.2018).

[102] W3C. *Ontology for Media Resources 1.0*. http://www.w3.org/TR/mediaont-10/. 2012, (accessed: 01.08.2018).

[103] Weibel, S., Kunze, J., Lagoze, C., and Wolf, M. *Dublin Core Metadata for Resource Discovery*. Tech. rep. 2070-1721. 1998.

[104] World Wide Web Consortium (W3C). *Resource Description Framework*. https://www.w3.org/RDF/. 2004, (accessed: 23.08.2018).

[105] World Wide Web Consortium (W3C). *SPARQL Query Language for RDF*. https://www.w3.org/TR/rdf-sparql-query/. 2008, (accessed: 28.08.2018).

[106] Wu, Z. and Palmer, M. "Verbs semantics and lexical selection". In: *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics. 1994, pp. 133–138.

[107] Zhang, J. "Evaluations on XML standards for actual applications". PhD thesis. University of British Columbia, 2008.

[108] Zhong, J., Zhu, H., Li, J., and Yu, Y. "Conceptual graph matching for semantic search". In: *International Conference on Conceptual Structures*. Springer. 2002, pp. 92–106.

[109] Zou, L., Özsu, M. T., Chen, L., Shen, X., Huang, R., and Zhao, D. "gStore: a graph-based SPARQL query engine". In: *The VLDB Journal—The International Journal on Very Large Data Bases* 23.4 (2014), pp. 565–590.