



HAL
open science

**ISEE : un système pour l'explication des événements
dans les environnements hybrides - mise en œuvre sur
des données de capteurs associées à des corpus
documentaires**

Nabila Guennouni

► **To cite this version:**

Nabila Guennouni. ISEE : un système pour l'explication des événements dans les environnements hybrides - mise en œuvre sur des données de capteurs associées à des corpus documentaires. Web. Université de Pau et des Pays de l'Adour, 2022. Français. NNT : 2022PAUU3034 . tel-04121008

HAL Id: tel-04121008

<https://theses.hal.science/tel-04121008>

Submitted on 7 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



ISEE : un système pour l'explication des événements dans les environnements hybrides - mise en oeuvre sur des données de capteurs associées à des corpus documentaires

Nabila GUENNOUNI

Rapporteurs :	Dr. Nathalie Aussenac-Gilles	Institut de Recherche en Informatique de Toulouse (IRIT), France
	PR. Chirine Ghedire	Laboratoire d'Informatique en Images et Systèmes d'Information (LIRIS), France
Examineurs :	PR. Max Chevalier	Institut de Recherche en Informatique de Toulouse (IRIT), France
	PR. Richard Chbeir	Univ. Pau & Pays Adour, France
Directeurs de thèse :	Dr. Christian Sallaberry	Univ. Pau & Pays Adour, France
	Dr. Sébastien Laborie	Univ. Pau & Pays Adour, France
Invité :	Dr. Elio Mansour	SAS Scient Analytics

Une thèse soumise dans le cadre des conditions
d'obtention du diplôme de docteur en informatique

Mai, 2022

Remerciement

Je voudrais avant tout remercier DIEU, le tout puissant, pour tous ses bienfaits trop souvent négligés, sans lui rien de tout cela ne serait possible.

Je tiens tout d'abord à remercier profondément mes directeurs de thèse, le Dr. Christian Sallaberry et le Dr. Sébastien Laborie pour leur confiance, pour la liberté qu'ils m'ont accordée et pour la qualité de leur encadrement. Je suis extrêmement reconnaissante au Dr. Christian Sallaberry pour sa supervision attentive, sa patience et ses conseils. Malgré ses responsabilités, il a su être présent quand il le fallait. Je suis également très reconnaissante au Dr. Sébastien Laborie pour son soutien continu, sa motivation, ses suggestions pratiques et ses commentaires constructifs pendant la réalisation de mon doctorat. Je voudrais aussi les remercier pour leurs qualités personnelles et humaines qui ont aussi beaucoup contribué à la réalisation de ce travail.

Je suis reconnaissante d'avoir fait partie du laboratoire LIUPPA. Je dois ma gratitude à "E2S" l'organisme de financement du projet. Rien n'aurait été possible sans son soutien financier et son intérêt pour la promotion de la science. J'apprécie profondément l'aide et le soutien du Pr. Richard Chbeir et du Dr. Elio Mansour, qui ont toujours apporté de bonnes idées lors des discussions du projet.

Je tiens à remercier l'IUT de Bayonne et du Pays Basque à Anglet pour m'avoir permis d'utiliser les locaux et le matériel. Je reconnais le grand soutien moral et émotionnel apporté par Hernán Humberto Álvarez Valera, Sabri Allani et Faisal Shahzad. Ces personnes étaient à mes côtés à chaque fois que j'en avais besoin. Je ne voudrais pas non plus oublier mes amis du Maroc. Je tiens à remercier tout particulièrement Amal El Kaid, Amine Boussik et Inas Benjelloun qui m'ont toujours encouragé à faire des études doctorales à l'étranger. Leur soutien et leur motivation m'ont permis d'atteindre cet objectif.

Je voudrais remercier mes parents qui m'ont toujours épaulé pendant ces années. Ils ont toujours cru en moi, ont fait preuve d'un soutien indéfectible dans cette expérience mais aussi tout au long de ma scolarité et de ma vie professionnelle. Merci à vous mes parents! sans vous, jamais je n'aurais pu y arriver.

Enfin, je remercie mon cher époux El Mahdi Dichaoui pour son soutien quotidien,

et son enthousiasme contagieux à l'égard de mes travaux comme de la vie en général. Son amour ne m'a procuré que confiance et stabilité. Je remercie le bon dieu qui a croisé nos chemins. Puisse le bon dieu nous procurer santé et longue vie.

Pour tout cela merci mon coeur!.

À toi, je dédicace ce travail, avec tout mon amour.

Résumé

En raison de leur potentiel pour l'amélioration de la sécurité, du confort, de la productivité et des économies d'énergie, les environnements connectés sont devenus omniprésents dans notre vie quotidienne. Ils ont eu un impact sur différents secteurs d'activité, tels que par exemple, les hôpitaux, les centres commerciaux, les fermes et les véhicules.

Afin d'améliorer encore plus la qualité de vie dans ces environnements, beaucoup d'applications proposant des services basés sur l'exploitation des données collectées par les capteurs ont vu le jour. La détection d'événements est l'un de ces services (*par exemple, la détection d'incendie, la détection des accidents vasculaires cérébraux pour les patients, la détection de la pollution atmosphérique*).

Généralement, quand un événement est déclenché dans un environnement connecté, la réaction naturelle d'un responsable est d'essayer de comprendre ce qui s'est passé et pourquoi l'évènement s'est déclenché. Pour trouver des réponses à ces questions, l'approche traditionnelle consiste à interroger manuellement les différentes sources de données à disposition (*le système d'information du réseau de capteurs et le système d'information gérant les corpus de documents*), ce qui peut s'avérer très fastidieux, très coûteux en matière de temps et nécessite un énorme effort de compilation.

Cette thèse s'intéresse à l'explication des événements détectés dans les environnements connectés, et plus précisément à ceux qui se produisent dans des environnements disposant de systèmes d'information (SI) hybrides (SI du corpus de documents et SI des données de capteurs de capteurs). Nous proposons un système intitulé ISEE (Information System for Event Explanation). ISEE est basé sur : (i) un modèle multidimensionnel pour la définition des événements dans les environnements connectés : ce dernier permet aux utilisateurs des environnements connectés de définir les événements selon différents axes de description (document et réseau de capteurs) favorisant ainsi le rapprochement entre les différentes sources de données; (ii) un modèle 5W1H adapté pour la structuration des explications d'événements : ce modèle s'inspire de l'approche 5W1H (*What, Who, When, Where, How et Why*) et l'adapte à notre contexte pour présenter aux utilisateurs des explications simples est faciles à comprendre; (iii) un processus pour l'interconnexion et le filtrage ciblé des ontologies de domaine : ce processus a pour objectif d'analyser l'ensemble des ontologies de domaine et de construire des explications aux évènements déclenchés dans les environnements connectés en s'appuyant sur des interconnexions sensibles au contexte (l'explication des événements); et (iv) un processus pour le classement des explications : ce processus a pour objectif de

classer par ordre de pertinence les explications construites par l'étape précédente en s'appuyant sur une métrique originale.

Nous proposons une solution générique qui peut être appliquée dans différents domaines. Néanmoins, trois expérimentations ont été conduites pour valider cette proposition dans le contexte d'un bâtiment et d'un parking connectés.

Table des matières

1	Introduction	1
1.1	L'évolution numérique	1
1.1.1	Les environnements connectés	1
1.1.2	La digitalisation des entreprises	2
1.2	Contexte de la thèse	4
1.2.1	Objectif de la thèse	5
1.2.2	Scénarios de motivation	5
1.2.3	Défis et hypothèses de travail	10
1.2.4	Verrous scientifiques	11
1.3	Contributions	11
1.3.1	Contribution 1. Modélisation des évènements dans les environnements connectés hybrides	12
1.3.2	Contribution 2. Interconnexion et filtrage ciblés des ontologies	13
1.4	Publications	14
1.5	Organisation du manuscrit	15
2	État de l'art	16
2.1	Introduction	16
2.2	La recherche d'information dans les environnements connectés	17
2.2.1	La recherche d'information classique dans les environnements connectés	18
2.2.2	La recherche d'information sémantique dans les environnements connectés	20
2.2.3	Synthèse et discussion	25
2.3	La recherche d'information dans les corpus documentaires	26
2.3.1	La recherche d'information classique dans les corpus documentaires	28
2.3.2	La recherche d'information sémantique dans les corpus documentaires	29
2.3.3	Systèmes question-réponse	32
2.3.4	Les approches 5W1H	35
2.3.5	Évaluation des systèmes de recherche d'information	37
2.3.6	Synthèse et discussion	38
2.4	Interconnexion de données issues d'environnements hybrides	41

2.4.1	Alignement d'ontologies	42
2.4.2	Liaison de données	46
2.4.3	Stratégies d'interconnexion	47
2.4.4	Synthèse et discussion	48
2.5	Bilan	49
3	Le système ISEE	52
3.1	Introduction	52
3.2	Vue d'ensemble	53
3.3	Définition des composantes d'un environnement hybride	56
3.4	Modélisation des évènements dans les environnements hybrides	58
3.4.1	Un modèle multidimensionnel pour la définition d'évènements dans les environnements connectés	59
3.4.2	Un modèle pour la définition des explications des évènements	62
3.5	L'explication d'évènements déclenchés dans un environnement hybride	66
3.5.1	Processus d'intégration des connaissances du domaine	67
3.5.2	Évaluation de l'alignement des ontologies	70
3.5.3	Le processus ISEE : Interconnexion et filtrage ciblés des ontologies de domaine	72
3.6	Conclusion	91
4	Expérimentation et Évaluation	93
4.1	Introduction	93
4.2	La plateforme ISEEapp	94
4.3	Évaluation du modèle de définition des évènements	97
4.3.1	Protocole expérimental	97
4.3.2	Résultats	100
4.4	Évaluation des résultats retournés par le processus d'explication des évènements	101
4.4.1	Preuve de concept : données simulées et exécution manuelle	102
4.4.2	Expérimentation sur la plateforme ISEEapp : Données réelles et exécution automatique	108
4.5	Évaluation de l'interface utilisateur pour l'explication d'évènements	113
4.5.1	Protocole expérimental	113
4.5.2	Résultat	114
4.6	Conclusion	116
5	Conclusion	118
5.1	Récapitulatif	118
5.2	Perspectives	120
A	ISEEapp	125
A.1	Introduction	125
A.2	Instanciation des ontologies	126

TABLE DES MATIÈRES

A.3 Alignement des ontologies	127
A.4 Évaluation des alignements	128
A.5 Définition des événements	129
A.6 Détection des événements	130
A.7 Explication d'événements	131
A.8 Conclusion	131
Bibliographie	134

Table des figures

1.1	Valeur des investissements dans les bâtiments intelligents par région du monde entre 2017 et 2025 en milliards de dollars US ¹	2
1.2	Dépenses mondiales consacrées à la digitalisation entre 2017 et 2025 ² . .	3
1.3	Le grand bâtiment de recherche ³	6
1.4	Recherche d'explication pour l'événement gaspillage de lumière	7
1.5	Le parking connecté	7
1.6	Recherche d'explication pour l'événement haut niveau de CO2	8
1.7	Modélisation de l'environnement en utilisant des ontologies de domaine	12
1.8	Définition de l'événement <i>gaspillage de lumière</i> et extension de l'ontologie réseau de capteurs	13
1.9	Processus d'interconnexion et de filtrage ciblés des ontologies de domaine	14
2.1	Schémas de flux de capteurs hétérogènes	21
2.2	Le processus de la RIS pour les réseaux de capteurs	21
2.3	Le processus classique des systèmes de la RI pour les corpus documentaires ⁴	27
2.4	Processus général des systèmes question-réponse ⁵	33
2.5	Le processus d'alignement ⁶	42
2.6	Fragments des ontologies réseaux de capteurs et automobile	43
2.7	Synthèse des techniques d'alignement d'ontologies ⁷	44
2.8	La liaison de données et l'alignement d'ontologies ⁸	46
3.1	Vue d'ensemble du système ISEE	53
3.2	Le modèle ISEE	67
3.3	Un exemple d'alignement d'ontologies	71
3.4	Un exemple de connexions $r_{C_{what}}$ et $r_{C_{who}}$ de concepts	77
3.5	Un exemple d'explication standard de l'événement gaspillage de lumière	80
3.6	Un exemple des connexions 4W des instances	85
3.7	Un exemple d'explication complète de l'événement gaspillage de lumière	88
4.1	Architecture de la plateforme ISEEapp	95
4.2	Évolution de la connectivité en fonction du nombre d'événements défini pour les trois ontologies EDOHE, LODE et SEM	101
4.3	Fiche d'information de l'employé Philippe de la Charrier	109
4.4	Capture d'écran du questionnaire d'évaluation de facilité de compréhension	114

TABLE DES FIGURES

4.5	Évaluation de la facilité de compréhension sur le 1 ^{er} exemple de l'évènement gaspillage de lumière	115
4.6	Évaluation de la facilité de compréhension sur le 2 ^{ème} exemple de l'évènement gaspillage de lumière	115
5.1	Le système ISEE et les perspectives	121
A.1	Capture d'écran d'un fichier d'alignement retourné par l'outil Agreement-MakerLight	127
A.2	Évaluation d'un alignement d'ontologie sur la plateforme ISEEapp	128
A.3	Choix des ontologies avant la définition d'un événement	129
A.4	Chargement d'une nouvelle ontologie par l'utilisateur dans la plateforme ISEEapp	129
A.5	Définition d'un événement dans la plateforme ISEEapp	130
A.6	Exemple d'explication d'un évènement gaspillage de lumière dans la plateforme ISEEapp	132

Liste des tableaux

2.1	Évaluation des ontologies réseau de capteurs	25
2.2	Techniques adoptées par les approches de la RI classique, de la RIS sémantique et des SQR pour les corpus documentaires	39
2.3	Techniques adoptées par les approches d'interconnexion de données	48
4.1	Observations collectées par les 6 types de capteurs	103
4.2	Résultat de l'étape 1 du processus ISEE	106
4.3	Résultat de l'étape 2 du processus ISEE	106
4.4	Résultat de l'étape 3 du processus ISEE	106
4.5	Évaluation des étapes 1 et 2 du système ISEE sur la base des trois paramètres d'évaluation suivants : précision, rappel et F1-score.	107
4.6	Évaluation de l'étape 3 du processus ISEE en utilisant la métrique Precision@n	107
4.7	Observations collectées par les 5 types de capteurs	109
4.8	Nombre de fiches d'information générées par poste	110
4.9	Évaluation de la complétude et de la cohérence	112

Liste des abréviations

5W1H	What, Who, When, Where, How et Why
AA	Apprentissage Automatique
ASL	Analyse Sémantique Latent
CD	Corpus Documentaire
CE	Complex Event
DCG	Discounted Cumulative Gain
EC	Environnement Connecté
EDOHE	Event Description Ontology for Heterogeneous Environments
EH	Environnement Hybride
HSSN	Hybrid Semantic Sensor Network
IDC	International Data Corporation
IoT	Internet of Things
IRI	Internationalized Resource Identifier
ISEE	Information System for Event Explanation
MSSN-onto	Multimedia SSN ontology
MRR	Mean Reciprocal Rank
NLP	Natural Language Processing
NoSQL	Not Only SQL
LODE	Linking Open Descriptions of Events
RDF	Resource Description Framework
RI	Recherche d'Information
RIS	Recherche d'Information Sémantique
SAREF4BLDG	SAREF extension For Building
SEM	Simple Event Model
SI	Système d'Information
SGBD	Systèmes de Gestion de Bases de Données Relationnelles
SGBDNoSQL	Systèmes de Gestion de Bases de Données Relationnelles NoSQL

SGBDST	S ystèmes de G estion de B ases de D onnées de S éries T emporelles
SPARQL	S imple P rotocol A nd R DF Q uery L anguage
SQL	S tructured Q uery L anguage
SQR	S ystème Q uestion- R éponse
SSN	S emantic S ensor N etwork
SUMO	S uggested U pper M erged O ntology
SVM	S upport- V ector M achine
TIC	T echnologies de l' I nformation et de la C ommunication
URL	U niform R esource L ocator
W3C	W orld W ide W eb C onsortium

Chapitre 1

Introduction

Dans ce premier chapitre, nous présentons tout d'abord les domaines d'activité liés à la thèse dans la section 1.1. La section 1.2 décrit précisément le contexte de la thèse (objectif, hypothèses de travail, verrous scientifiques, etc.). Nous présentons également dans cette section deux scénarios qui illustrent la motivation de ce travail et les défis à considérer. Ensuite, nous présentons brièvement notre proposition et les contributions qui en découlent dans la section 1.3. Les publications qui valident ces contributions sont présentées dans la section 1.4. Enfin, la section 1.5 présente l'organisation du reste du manuscrit.

1.1 L'évolution numérique

Dans cette section, nous introduisons brièvement les deux domaines d'activité liés à la thèse. Nous décrivons tout d'abord **l'évolution des environnements connectés** de nos jours et nous présentons quelques statistiques récentes dans la section 1.1.1. Ensuite, de la même façon, nous décrivons le concept de **digitalisation des entreprises** et nous présentons quelques statistiques dans la section 1.1.2.

1.1.1 Les environnements connectés

Ces dernières années ont connu un intérêt accru pour les environnements connectés (EC) [68]. L'augmentation du coût de l'énergie et les préoccupations environnementales croissantes, telles que les émissions de carbone et la pollution, obligent les constructeurs à s'orienter de plus en plus vers des infrastructures de bâtiments intelligents. Généralement, un environnement connecté peut être défini comme une infrastructure qui héberge un ensemble de capteurs travaillant en collaboration pour fournir des données à diverses applications (par exemple, le contrôle à distance des appareils électriques, la gestion de la sécurité et de la consommation d'énergie).

Les environnements connectés sont devenus omniprésents dans notre vie quotidienne, ils ont eu un impact sur différents secteurs, tels que par exemple, les hôpitaux, les centres commerciaux, les fermes, les bâtiments et les véhicules. Les investissements

dans le marché des environnements connectés n’ont pas cessé de croître à une cadence très rapide.

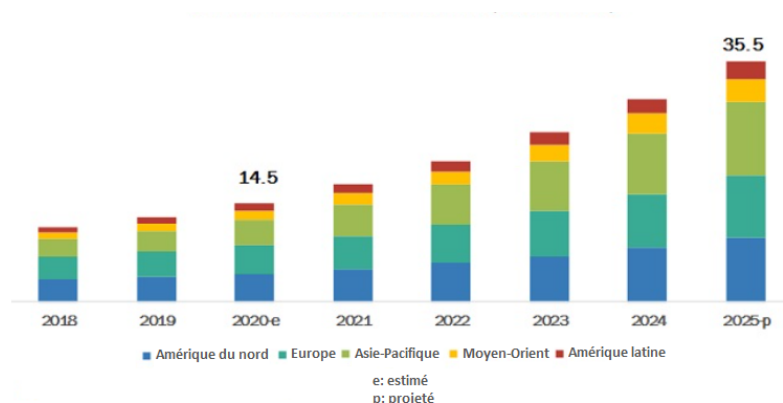


FIGURE 1.1 – Valeur des investissements dans les bâtiments intelligents par région du monde entre 2017 et 2025 en milliards de dollars US ¹

La figure 1.1 montre la répartition des investissements dans les bâtiments intelligents par région du monde entre 2017 et 2025. On peut voir que la valeur des investissements connaît une importante augmentation (de 14.5 milliards de dollars en 2020 à 35.5 milliards de dollars projetés en 2025). De plus, malgré l’impact économique de la pandémie COVID-19, l’intérêt du côté des consommateurs a également évolué. Une nouvelle mise à jour du média américain IDC (International Data Corporation) ² montre que les dépenses liées aux EC ont augmenté de 8,2% sur un an pour atteindre 742 milliards de dollars en 2020, contre 14,9% de croissance prévue dans la version de novembre 2019. L’évolution de l’intérêt du côté des consommateurs, peut-être expliquée par le fait que le déploiement des technologies liées aux EC, leurs permet de profiter de plusieurs services et applications intelligents [169]. Un exemple typique de ces services auquel nous nous intéressons dans cette thèse est **la détection des événements** (par exemple, détection d’incendie, détection d’intrusion, détection d’accident routier, etc.). L’IDC s’attend à ce que les dépenses mondiales liées aux EC reviennent à des taux de croissance à deux chiffres en 2021 et atteignent un taux de croissance annuel composé de 11,3% au cours de la période de prévision 2020-2024.

1.1.2 La digitalisation des entreprises

L’émergence des technologies de l’information et de la communication (TIC) qui a commencé vers 1980 avec le développement du réseau Internet [82] et puis avec les appareils mobiles, les réseaux sociaux, le Big Data et le web sémantique, a révolutionné les pratiques de travail et de communication dans le monde. Dans le passé, lorsqu’une

1. Source : MarketsandMarkets Analysis, Décembre 2021, <https://www.marketsandmarkets.com/Market-Reports/retail-iot-market-43188550.html>

2. Source : IDC, Décembre 2021, <https://www.idc.com/getdoc.jsp?containerId=prUS46609320>

entreprise recevait une commande d'un client, par exemple un service demandé ou un produit acheté, la commande devait passer par un processus "papier" qui était transmis à différents services d'employé à employé. Tout au long de ce processus, la commande devait souvent être ressaisie lors de son passage dans les différents services, ce qui était coûteux en terme de temps et augmentait le risque d'erreur. Aujourd'hui, il est devenu indispensable pour les entreprises de faire évoluer leurs stratégies, de s'adapter à leurs environnements et au monde qui les entoure. **La digitalisation des entreprises** consiste à prendre un virage numérique dans la stratégie et l'organisation interne de l'entreprise [46]. En effet, les entreprises peuvent tirer de multiples bénéfices de la digitalisation [111, 107], tels que :

- Un gain de temps important à travers l'automatisation des processus, par exemple, le suivi automatique des factures impayées;
- Une communication plus rapide, plus simple et moins coûteuse entre les différents départements de l'entreprise, par exemple, bureau et chantier;
- Un accès rapide aux informations grâce aux sauvegardes sur des serveurs décentralisés (Cloud), ce qui permet de travailler à distance quoi qu'il arrive à l'entreprise (inondation, incendie, etc.);
- Une meilleure connaissance des clients grâce aux données collectées ce qui permet d'adapter les offres en fonction des besoins et des attentes.

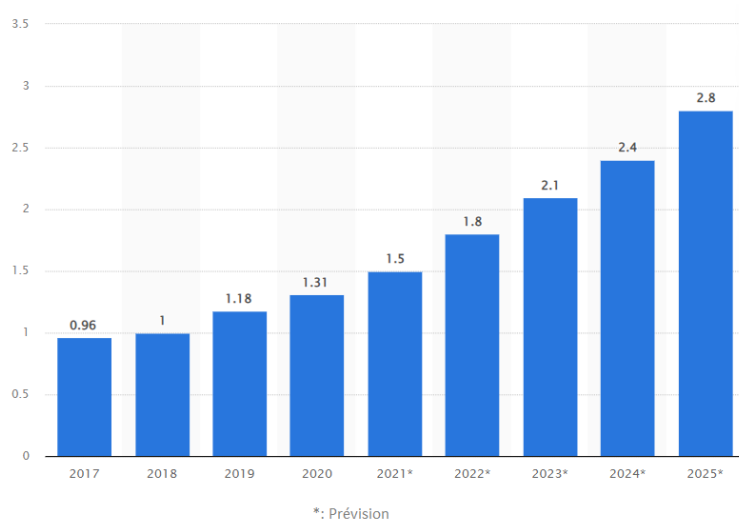


FIGURE 1.2 – Dépenses mondiales consacrées à la digitalisation entre 2017 et 2025³

La figure 1.2 met en évidence les augmentations des dépenses mondiales consacrées à la digitalisation. Les dépenses sont passées de 0.96 trillions de dollars en 2017 à 2.8 trillions de dollars projetés en 2025. La vitesse de cette transformation est régie

3. Source : Statista, Décembre 2021, <https://www.statista.com/statistics/870924/worldwide-digital-transformation-market-size/>

par les progrès de la technologie de connectivité, les changements de comportement des consommateurs, l'émergence de nouveaux modèles commerciaux, mais aussi, le contexte actuel de la pandémie COVID-19 où l'économie mondiale subit des crises sans précédent. La digitalisation est devenue la clé de nombreuses entreprises pour s'adapter et surmonter la situation actuelle.

Une des étapes capitales de la digitalisation à laquelle nous nous intéressons dans cette thèse, est la gestion des données de l'entreprise de nature hétérogène par un système d'information (SI). Le système d'information a un rôle central dans le fonctionnement de l'entreprise, il permet de créer, collecter, stocker, traiter et modifier des informations sous divers formats (par exemple, fichiers texte, fichiers multimédias, tableurs, etc.) [112]. Par exemple, une entreprise utilise son SI pour traiter ses comptes financiers, gérer ses ressources humaines et atteindre des clients potentiels.

Pour conclure, dans cette première partie du chapitre introduction, nous nous sommes intéressés à l'évolution numérique dans les environnements connectés et ensuite dans les entreprises. En ce qui concerne les environnements connectés, nous avons vu que l'intérêt pour ces technologies ne cesse d'augmenter que ce soit au niveau des grandes industries ou au niveau des consommateurs. Concernant la digitalisation des entreprises, nous avons vu que ce n'est plus un choix mais une obligation compte-tenu du contexte sanitaire actuel et de différents autres facteurs.

1.2 Contexte de la thèse

Au cours de la dernière décennie, afin d'exploiter le potentiel des EC, un nombre croissant de travaux de recherche ont été menés pour proposer des services supplémentaires basés sur l'utilisation des données collectées par les capteurs [169]. La détection d'événements est l'un de ces services. Les événements peuvent être définis comme "*quelque chose qui se passe quelque part*"⁴. Dans la littérature deux types d'événements sont distingués [58] :

1. **Les faits** : ils correspondent à des informations d'actualité, à des événements historiques ou à des phénomènes observés dans le monde physique, par exemple, l'élection du président Joe Biden, le déclenchement de la première guerre mondiale, le déclenchement d'un incendie dans un bâtiment, etc. ;
2. **Les événements socio-culturels** : ils sont des événements planifiés et auxquels un public est associé, par exemple, les concerts, les rencontres sportives, les festivals, etc.

Cette thèse s'intéresse à l'explication des faits (événements de la catégorie 1.), et plus particulièrement ceux qui se produisent dans des environnements connectés impliquant des SI hétérogènes (SI de corpus de documents et SI de réseau de capteurs). Nous appelons dans le reste de cette thèse ces environnements exploitant les SI de corpus de documents et les SI de réseau de capteurs **les environnements hybrides**.

4. <http://www.larousse.fr/dictionnaires/francais/événement/31839>

1.2.1 Objectif de la thèse

La thèse considère les environnements connectés associés à des corpus documentaires. L'importance des corpus documentaires réside dans le fait qu'ils constituent une source d'information essentielle pour compléter les données des capteurs. Bien entendu, il peut exister des exemples d'environnements connectés associés à des bases de données contenant des informations sur ces environnements (objets installés dans l'environnement, informations sur les employés et sur les clients, etc.). Dans ce cas de figure, l'exploitation de ces informations est beaucoup plus simple par rapport à celle des corpus documentaires. Néanmoins, dans le cadre de cette thèse, nous avons décidé de rester général en supposant que les informations sur l'environnement sont présentes dans des corpus documentaires.

L'objectif de cette thèse est d'aider les utilisateurs des environnements connectés (par exemple, les propriétaires de maisons intelligentes, les gestionnaires de bâtiments intelligents) à comprendre pourquoi un certain événement a eu lieu. L'objectif en général, est de fournir un système qui permet aux utilisateurs de trouver facilement des pistes d'explication aux événements déclenchés. Dans un premier temps, le système doit permettre aux utilisateurs de définir les événements qu'ils souhaitent détecter (par exemple, haut niveau de CO₂, gaspillage de lumière). Ensuite, les données étant représentées sous différents formats (par exemple, fichiers TXT, PDF, CSV, etc.) et dispersées dans plusieurs systèmes d'information (SI documents et SI capteurs), le système doit être capable d'interconnecter et d'exploiter ce nuage de données pour construire les explications. Enfin, les explications doivent être claires et facilement compréhensibles par n'importe quel type d'utilisateurs.

1.2.2 Scénarios de motivation

Pour illustrer les motivations de notre travail, nous proposons deux scénarios de motivation. Le premier scénario traite l'exemple de l'événement *gaspillage de lumière* dans un grand bâtiment de recherche (section 1.2.2.1). Le deuxième scénario de motivation porte sur l'exemple de l'événement *haut niveau de CO₂* dans un parking (section 1.2.2.2). Nous avons choisi de présenter ces deux scénarios de motivation pour montrer l'importance de la prise en compte du niveau d'urgence de l'événement et pour montrer la généralité du problème.

1.2.2.1 1^{er} scénario de motivation : Gaspillage de lumière dans un bâtiment de recherche

Nous considérons un exemple réel d'un grand bâtiment de recherche constitué de quatre étages (Figure 1.3). Divers capteurs (par exemple, capteurs de lumière, capteurs de température, capteurs de mouvement, etc.), entités mobiles (par exemple, personnes, ordinateurs portables, etc.) et entités statiques (par exemple, système d'extraction d'air, portes, lampes, etc.) sont déployés dans les différentes pièces/étages. Un système d'information associé gère également un corpus de documents impliquant une grande

quantité de données semi-structurées/non structurées, à savoir des documents ayant des contenus, des structures et des formats différents (par exemple, le site web de l'entreprise, le fichier de données des employés, les fiches techniques des capteurs et des ordinateurs portables, les rapports de maintenance, etc.).

Supposons maintenant qu'une consommation d'énergie anormale a été détectée dans le bureau 413 du quatrième étage la nuit dernière. Le responsable du bâtiment aimerait évidemment comprendre pourquoi cela s'est produit (par exemple, pourquoi l'événement a été déclenché? Quel employé est affecté au bureau n° 413? Qui était le veilleur de nuit hier? Y a-t-il des données intéressantes provenant d'autres capteurs? Des dysfonctionnements d'équipements ont-ils été signalés récemment au quatrième étage?...). La figure 1.4 présente la méthode classique de recherche d'explication d'évé-



FIGURE 1.3 – Le grand bâtiment de recherche⁵

nement par un responsable d'EC. Tout d'abord, en utilisant le système d'information du réseau de capteurs, le responsable est submergé par les données brutes des pièces (la partie gauche de la figure 1.4) : concentration de CO₂, humidité de l'air ambiant, température ambiante, luminosité et données du capteur de mouvement PIR. De même, les ressources documentaires correspondantes sont énormes et nombreuses, un processus d'interrogation efficace est difficile (la partie droite de la figure 1.4). La luminosité détectée par le capteur de luminosité dans le bureau n°413 à 20h correspond-elle à la visite du veilleur de nuit Garret? Perrin est-il l'employé qui a oublié la lumière allumée lorsqu'il a quitté son bureau à 17h? Ou bien est-ce que c'est la lampe du bureau 413 qui est défectueuse?

De plus, selon le contexte, le responsable du bâtiment peut avoir besoin d'une explication dans un délai plus ou moins court, par exemple, l'alarme du bâtiment sonne à chaque fois qu'il y a des lumières qui restent allumées la nuit, le responsable a donc besoin de savoir ce qui s'est passé en urgence, ou bien, il s'agit seulement d'une analyse suite à une prise de conscience énergétique et donc l'urgence est basse. Nous devons donc tenir compte de cette contrainte et c'est ce que nous allons présenter dans le second scénario qui suit (section 1.2.2.2).

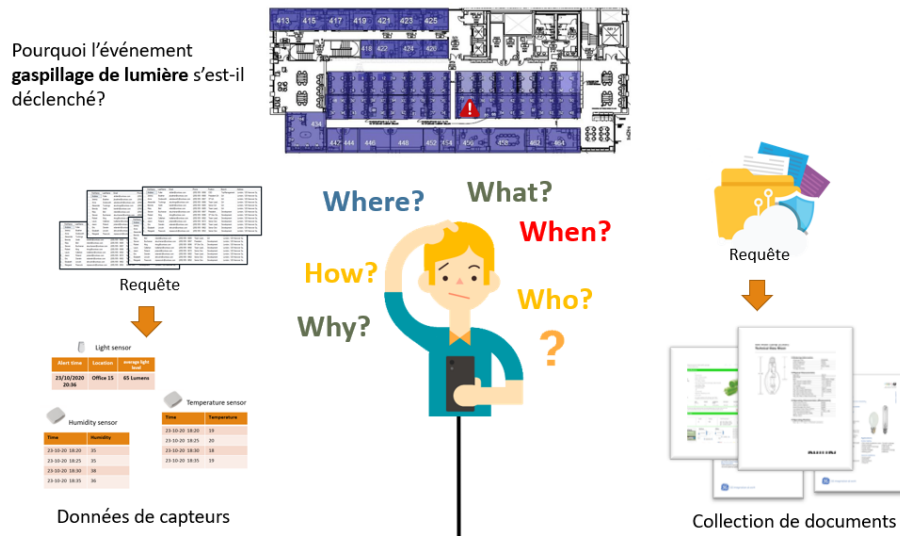


FIGURE 1.4 – Recherche d'explication pour l'événement gaspillage de lumière

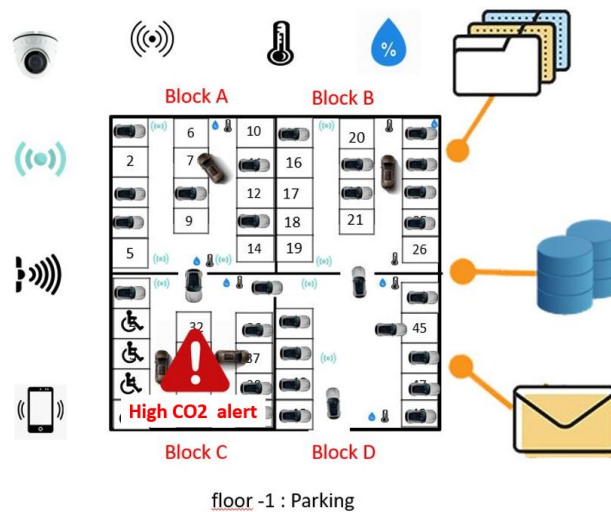


FIGURE 1.5 – Le parking connecté

1.2.2.2 2^{ème} scénario de motivation : Haut niveau de CO2 dans un parking connecté

La figure 1.5 présente un exemple de parking connecté comportant quatre blocs. Divers capteurs (par exemple, détecteur de fumée, capteur de CO2, capteur de température, etc.), entités mobiles (par exemple, voitures, personnes) et entités statiques (par exemple, système d'extraction d'air, portes, climatiseurs, lampes, etc.) sont déployés dans les différents blocs. Un corpus de documents est attaché au parking (par exemple, des fichiers des abonnés, des fiches techniques de capteurs et de voitures, des rapports

5. <https://citris-uc.org/about/sutardja-dai-hall/about-facilities/floorplans/>

de maintenance, etc.). Les données du réseau de capteurs et du corpus de documents sont gérées par deux SI différents.

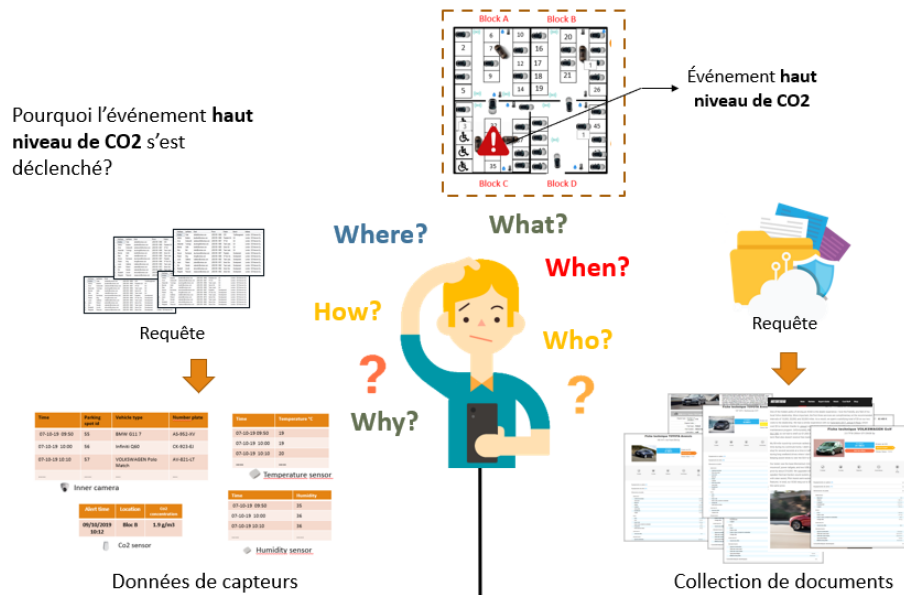


FIGURE 1.6 – Recherche d'explication pour l'événement haut niveau de CO2

Supposons maintenant qu'une alerte haut niveau de CO2 se produit dans le bloc C à 10h12. Un responsable de parking souhaiterait comprendre pourquoi cela s'est produit. La figure 1.6 présente la méthode classique de recherche d'explication en utilisant l'exemple de l'événement haut niveau de CO2.

En utilisant le système d'information du réseau de capteurs, le responsable obtiendra toutes les données brutes des différents capteurs du bloc C (la partie gauche de la figure 1.6) :

- Capteur de CO2 : heure de l'alerte (09/10/2019 10 :12), niveau de CO2 dans le bloc C (1250ppm);
- Capteur de température : température du bloc C (09/10/2019 10 :10 19°C, 09/10/2019 10 :12 20°C, etc.);
- Caméra de surveillance : voitures circulant dans le bloc C au moment de l'alerte (BMW G11 7, Infiniti Q60, VOLKSWAGEN Polo Match, Renault TCe 120 EDC, etc.);
- Capteur d'humidité : taux d'humidité dans le bloc C (09/10/2019 10 :10 35, 09/10/2019 10 :12 36, etc.).

L'utilisateur est submergé par de nombreuses données mais sans pistes d'explication. De même, si le responsable tente d'obtenir des informations sur les émissions de CO2 par le biais du système d'information gérant les documents, il sera submergé par une grande quantité de données (la partie droite de la figure 1.6), par exemple, tous les documents contenant le mot clé CO2, les fiches techniques de tous les modèles de voitures enregistrés dans le système, les fiches techniques des capteurs de CO2, etc. Afin

de trouver une explication à l'événement, le responsable du parking doit donc faire des allers-retours entre les deux SI, par le biais de multiples requêtes. Voici un exemple de recherche possible afin d'illustrer ces allers-retours :

- Interroger le SI du réseau de capteurs pour obtenir l'heure et la localisation de l'alerte;
- Sur la base de l'heure et du lieu de l'alerte, interroger les données des caméras pour obtenir la liste des voitures du bloc C avant le déclenchement de l'événement;
- Interroger plusieurs fois le système d'information documentaire pour obtenir le niveau d'émission de CO₂ de chacune de ces voitures;
- Interroger les corpus de documents pour vérifier s'il y a eu des problèmes techniques ou des rapports de maintenance récents sur les équipements du bloc C.

Effectuer plusieurs requêtes sur deux systèmes différents pour rechercher l'explication d'un événement, demande beaucoup de temps et d'efforts. L'utilisateur doit avoir plusieurs expertises pour être capable d'interroger les deux SI. Il doit également pour chaque requête, consulter un grand nombre de documents pour extraire les informations pertinentes. Il doit ensuite compiler toutes ces données pour finalement tirer la conclusion suivante : les trois voitures "BMW G11 7", "Infiniti Q60" et "VOLKSWAGEN Polo", qui circulaient dans le bloc C peu de temps avant l'alerte, ont des caractéristiques très polluantes et peuvent être à l'origine de l'alerte. Enfin, dans le cas d'un niveau d'urgence élevé (par exemple, un incendie), il n'est pas opportun de procéder de la sorte.

1.2.2.3 Discussion

Lorsqu'un événement se déclenche dans un environnement hybride, le responsable de cet environnement rencontre des difficultés pour trouver une explication à cet événement. Ceci est dû aux facteurs suivants :

(i) **Portée des données** : un large spectre de données dispersées dans des sources de données hétérogènes avec des structures, des langages et des versions différentes (par exemple, des tableaux de données structurées dans le SI du réseau de capteurs et des fiches de données semi-structurées ou non structurées dans le SI des documents).

(ii) **Recherche manuelle** : les systèmes d'information existants pour les réseaux de capteurs ne permettent pas d'interconnecter le réseau de capteurs et le corpus de documents. C'est à l'utilisateur d'établir ces connexions par le biais de requêtes multiples : il doit d'abord interroger le réseau de capteurs, puis explorer les pistes intéressantes en interrogeant plusieurs fois les deux sources de données. L'utilisateur doit faire des allers-retours entre le réseau de capteurs et le corpus de documents. Il pourrait être submergé par une grande quantité de données inutiles. De plus, l'utilisateur doit reformuler ses requêtes de nombreuses fois pour affiner les résultats jusqu'à obtenir l'interprétation correcte de l'événement, ce qui est très fastidieux, prend beaucoup de temps et nécessite un énorme effort de compilation.

(iii) **Résultats mal structurés** : à la fin de ce processus, les éléments constituant l'interprétation sont dispersés dans divers documents et données de capteurs (par exemple, l'annuaire du personnel, les dossiers des employés affectés au bureau, le rapport d'accès au badge du bureau, la liste des équipements pouvant produire de la lumière dans le bureau). C'est à nouveau au responsable de l'environnement connecté de faire la synthèse.

1.2.3 Défis et hypothèses de travail

Quand un événement se déclenche, l'utilisateur souhaite savoir pourquoi ce dernier s'est déclenché et interroge le système à travers la requête "pourquoi l'événement s'est déclenché?".

Compte-tenu de tout ce que nous avons vu dans la section 1.2.2, les défis suivants se présentent :

— **Défi 1. Recherche d'information dans des ressources de données hétérogènes**

Le premier défi concerne l'interrogation automatisée de plusieurs ressources gérées par différents SI afin de collecter les données et ensuite construire la réponse à la requête de l'utilisateur. Ce défi se décline en deux sous-défis :

■ **Défi 1.1. Interrogation automatisée de différentes sources de données**

Comment interroger automatiquement des sources de données hétérogènes (différents langages et méthodes de requêtage ainsi que différents formats et structures de données), afin de collecter les informations nécessaires à la réponse de la requête de l'utilisateur ?

Pour résoudre ce défi, nous proposons l'hypothèse de travail suivante :

Hypothèse 1. Représentation sémantique

L'utilisation des représentations sémantiques (les ontologies), nous permettra de décrire les deux sources de données (documents et capteurs) en utilisant une représentation commune de la connaissance. Ainsi, nous pourrons les interroger conjointement en utilisant une même méthode de requêtage.

■ **Défi 1.2. Combinaison des résultats issus de requêtes multiples**

Une fois les SI interrogés, nous recevons des réponses sous différents formats (par exemple, des documents textes, des observations de capteurs, etc.), comment extraire les informations pertinentes de ces ressources et les combiner pour répondre à la requête de l'utilisateur ?

Nous proposons pour ce défi l'hypothèse de travail suivante :

Hypothèse 2. Alignement et l'approche 5W1H pour la recherche d'explication

L'utilisation des techniques d'alignements associée aux représentations sémantiques (hypothèse 1.) nous permettra dans un premier temps d'établir des **connexions classiques** entre les données de capteurs et les données du corpus de documents. Ces alignements étant génériques et déconnectés de

notre contexte (l'explication des événements), ils ne nous permettront pas d'établir toutes les connexions nécessaires à la construction des explications. Nous faisons donc l'hypothèse que des **connexions sensibles au contexte** inspirées des 5W1H (*What, Who, When, Where, How* et *Why*) seront nécessaires. Ces connexions nous permettront de guider le processus de recherche et de construire des explications pertinentes.

— **Défi 2. Présentation d'une réponse bien structurée**

Comment structurer la réponse finale de manière simple, claire et facile à comprendre?

Pour faire face à ce défi, nous proposons l'hypothèse suivante :

Hypothèse 3. Approche 5W1H pour la représentation de l'explication

La représentation des explications sous la forme de questions réponses 5W1H garantira un aspect simple et intuitif des explications construites par le système. Il faudra toutefois adapter cette représentation à notre contexte.

1.2.4 Verrous scientifiques

La thèse porte sur deux domaines de recherche principaux : la représentation des connaissances et la gestion des connaissances.

Concernant la représentation des connaissances, nous nous intéressons dans un premier temps à **la modélisation des événements dans des environnements hybrides**. Nous répondrons à la question suivante : comment définir un événement selon différents axes de description (capteurs et documents)?; Ensuite, nous abordons le problème de **la modélisation des explications des événements**. La question posée ici est la suivante : comment structurer les explications des événements d'une façon simple et facile à comprendre?

Quant à la gestion des connaissances, nous abordons le problème de **l'interconnexion ciblée des ontologies** que ce soit au niveau conceptuel ou au niveau des instances. Le mot "ciblée" fait référence à l'interconnexion des ontologies guidée par les données issues des événements. La question qui se pose ici est la suivante : comment bien exploiter les données issues des événements (définition et déclenchement) pour interconnecter efficacement l'ensemble des ontologies de domaine afin de produire des pistes d'explication?

1.3 Contributions

Comme expliqué précédemment (section 1.2), notre environnement consiste en deux SI : un SI réseau de capteurs et un SI corpus de documents. Comme l'indique notre *hypothèse 1* (section 1.2.3), nous supposons que nous disposons de plusieurs ontologies de domaine pour modéliser la sémantique du réseau de capteurs et du corpus de documents (ex. une ontologie ressources humaines et une ontologie bâtiment). Ces

ontologies sont ensuite instanciées en utilisant les données de capteurs et du corpus documentaire. Cette étape est présentée dans la figure 1.7.

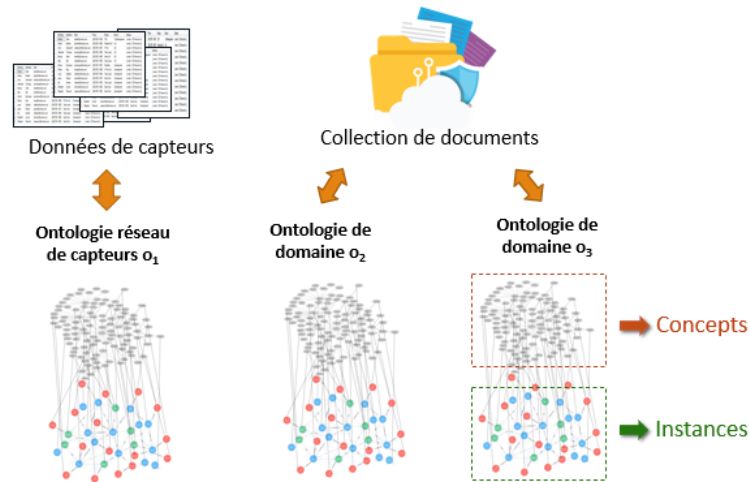


FIGURE 1.7 – Modélisation de l’environnement en utilisant des ontologies de domaine

Les contributions principales de la thèse se déclinent comme suit :

- **Contribution 1.** Modélisation des événements dans les environnements hybrides
- **Contribution 2.** Interconnexion et filtrage ciblés des ontologies

Dans ce qui suit nous détaillons chacune de ces deux contributions respectivement dans les sections 1.3.1 et 1.3.2.

1.3.1 Contribution 1. Modélisation des événements dans les environnements connectés hybrides

Cette première contribution concerne la modélisation d’un l’événement d’un point de vue (i) définition de ce dernier dans le système (label, critères et conditions du déclenchement, etc.) et (ii) structuration de l’explication retournée par le système. Cette contribution se décline donc en deux sous-contributions :

- **Contribution 1.1. Un modèle multidimensionnel pour la définition des événements dans les environnements connectés**

Ce modèle est utilisé pour étendre l’ontologie réseau de capteurs avec la notion d’événement. Il permet aux utilisateurs de définir un événement sous forme de plusieurs dimensions, par exemple, *Time*, *Location* et *Features*. La partie gauche de la figure 1.8 montre un exemple de définition de l’événement *gaspillage de lumière* précédemment mentionné dans la section 1.2.2.1, tandis qu’au centre de la figure, nous pouvons voir que l’ontologie réseau de capteurs a été étendue avec cette définition.

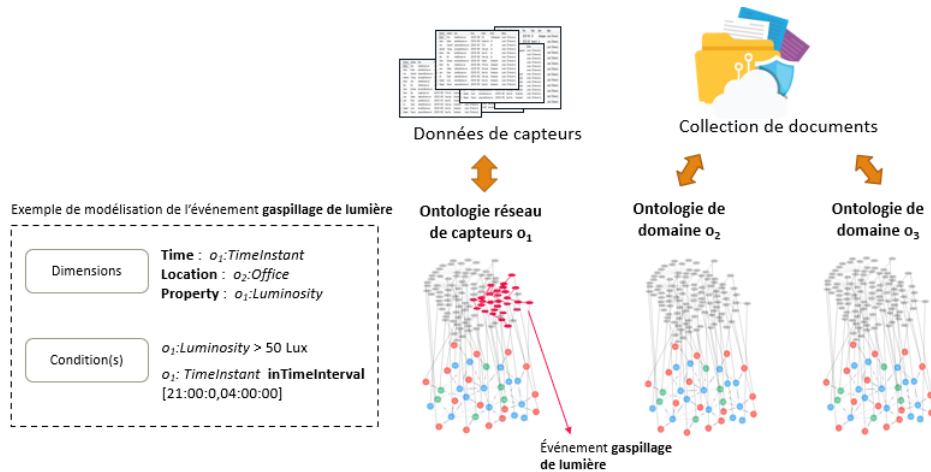


FIGURE 1.8 – Définition de l'événement *gaspillage de lumière* et extension de l'ontologie réseau de capteurs

L'utilisateur décrit ces dimensions en utilisant **des concepts des ontologies de domaine et des contraintes**, par exemple, pour décrire la dimension *Location*, l'utilisateur peut utiliser le concept $o_2:Office$ issu de l'ontologie bâtiment pour indiquer au système que l'événement peut être détecté dans un bureau du bâtiment. L'utilisation des concepts des ontologies de domaine pour décrire les dimensions des événements, permet de les relier aux données du corpus documentaire qui a servi à instancier ces ontologies. L'utilisateur peut également associer **des contraintes** à ces dimensions, par exemple, pour la dimension *Time* décrite en utilisant le concept $o_1:TimeInstant$, l'utilisateur peut définir la contrainte ($o_1:TimeInstant \text{ inTimeInterval } [21:00:00, 04:00:00]$) pour que l'événement ne soit détecté que durant l'intervalle $[21:00:00, 04:00:00]$.

- **Contribution 1.2. Un modèle pour la définition des explications des événements**
Ce modèle s'inspire de l'approche 5W1H [72, 189] utilisée dans la littérature dans les systèmes question-réponse pour structurer les explications construites par le système. Nous attribuons à chaque élément 5W1H (*What, Who, When, Where, How* et *Why*) une signification adaptée au contexte d'explication des événements dans les EC, par exemple, *Who* fait référence au capteur qui a détecté l'événement, *Why* décrit les différentes pistes d'explication de l'événement.

1.3.2 Contribution 2. Interconnexion et filtrage ciblés des ontologies

Nous proposons un processus pour l'explication des événements détectés dans les environnements hybrides. Ce processus se décline sous deux sous contributions :

- **Contribution 2.1. Un processus pour l'interconnexion et le filtrage ciblés des ontologies de domaine**

Ce dernier se base sur les données issues de la définition de l'événement (contribution 1.1.) pour établir des interconnexions sensibles au contexte entre les concepts des ontologies de domaine (interconnexion des concepts, figure 1.9). Ensuite, lorsque l'événement se déclenche (exemple des données issues du déclenchement de l'évènement gaspillage de lumière, figure 1.9), en se basant sur les informations nouvellement acquises et sur les interconnexions construites précédemment, le processus établit un filtrage des concepts et ensuite des instances pour ne garder que celles potentiellement reliées au déclenchement de l'événement en question (filtrage des concepts, filtrage des instances, figure 1.9).

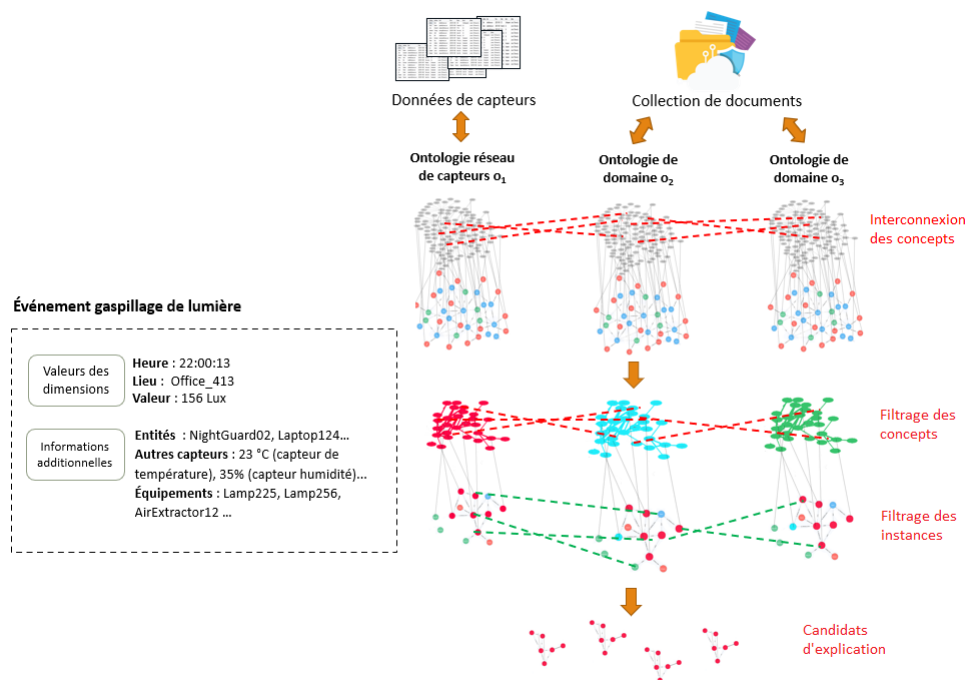


FIGURE 1.9 – Processus d'interconnexion et de filtrage ciblés des ontologies de domaine

— **Contribution 2.2. Un processus pour le classement des explications**

Le rôle de ce processus est d'analyser l'ensemble des instances filtrées (contribution 2.1.), de construire des candidats d'explication et les classer par ordre de pertinence (candidats d'explication, figure 1.9)

1.4 Publications

Les contributions de la thèse sont publiées et soumises aux conférences suivantes :

— **Présentation de la problématique et des hypothèses de travail**

Nabila Guennouni : Interprétation d'événement dans un système d'information hétérogène. INFORSID 2020 JCJC : 21-24

— **Présentation de la contribution 1.**

Nabila Guennouni, Christian Sallaberry, Sébastien Laborie, Richard Chbeir : A Novel Framework for Event Interpretation in a Heterogeneous Information System. MEDES 2020 : 140-148

— **Présentation de la contribution 2.**

Nabila Guennouni, Christian Sallaberry, Sébastien Laborie, Richard Chbeir, Elio Mansour : ISEE : A heterogeneous information system for event explainability in smart connected environments. Internet of Things 2021 : 100457

— **Présentation de la plateforme développée**

Nabila Guennouni, Christian Sallaberry, Sébastien Laborie, Richard Chbeir : ISEEapp : An Event Explanation Prototype bridging the gap between sensor network and document corpora data. IE 2022 - Accepted

1.5 Organisation du manuscrit

Le reste de la thèse est organisé comme suit :

Chapitre 2 Ce chapitre est dédié à l'état de l'art. Nous faisons un tour d'horizon des principaux axes de recherche explorés dans le cadre de la thèse. Nous dressons tout d'abord, un état de l'art des travaux concernant la recherche d'information classique et sémantique dans les EC. Ensuite, de la même façon, nous faisons le point sur la recherche d'information classique et sémantique dans les corpus documentaires. Enfin, nous parcourons les travaux concernant l'interconnexion des données issues des environnements hybrides.

Chapitre 3 Ce chapitre présente notre proposition, le système ISEE (Information System for Event Explanation) pour l'explication des événements dans les environnements hybrides. Nous présentons tout d'abord une vue d'ensemble du système ISEE. Ensuite, nous détaillons chacune de ses composantes en faisant l'accent sur les parties qui constituent des contributions de cette thèse. Enfin, nous donnons quelques exemples des résultats retournés par le système en s'appuyant sur le scénario de l'événement gaspillage de lumière (section 1.2.2.1).

Chapitre 4 Ce chapitre présente brièvement la plateforme ISEEapp qui concrétise les contributions du chapitre 3 sous forme d'une application Web. Ensuite, une évaluation expérimentale de ces contributions est détaillée. Cette évaluation est composée de trois expérimentations qui sont conduites dans deux contextes différents, un bâtiment et un parking connectés. La première expérimentation évalue le modèle de définition d'événements (Contribution 1.). La deuxième expérimentation évalue le processus d'explication des événements (Contribution 2.). Enfin la troisième expérimentation évalue l'interface utilisateur pour l'explication des événements (Contributions 1. et 2.). Les expérimentations montrent des résultats prometteurs.

Chapitre 5 Ce chapitre conclut le manuscrit par une récapitulation de tous les chapitres susmentionnés et discute en détail les prochaines étapes et les orientations potentielles des recherches futures.

Chapitre 2

État de l'art

2.1 Introduction

Dans le chapitre 1, nous avons présenté l'objectif de cette thèse **l'explication des événements dans les environnements connectés basée sur le recoupement de données issues de capteurs et de corpus documentaires**. Communément, pour trouver une explication à un événement déclenché dans un environnement hétérogène, l'utilisateur doit (i) interroger à travers une ou plusieurs requêtes le système d'information des données de capteurs; (ii) interroger à travers une ou plusieurs requêtes le corpus documentaire; (iii) combiner les informations issues des deux sources de données et bâtir le sous-ensemble d'informations pertinentes (l'explication de l'événement). Par conséquent, dans ce chapitre, nous présentons l'état de l'art autour de trois axes : la recherche d'information (RI) dans les environnements connectés (section 2.2), la RI dans les corpus documentaires (section 2.3) et l'interconnexion de données hétérogènes (section 2.4).

Dans le premier axe, nous définissons tout d'abord formellement ce qu'est un environnement connecté. Ensuite, nous détaillons comment les systèmes de recherche classiques pour les réseaux de capteurs fonctionnent et comment les systèmes de recherche sémantique apportent des améliorations par rapport aux systèmes classiques.

Dans le deuxième axe, nous nous consacrons aux corpus documentaires. Nous présentons diverses approches de la RI classique et de la RI sémantique pour les corpus documentaires. Ensuite, nous nous intéressons aux systèmes question-réponse et nous mettons l'accent sur la technique 5W1H, puisque nous l'exploiterons dans nos contributions.

Le troisième axe se focalise sur l'interconnexion de données hétérogènes. Nous expliquons tout d'abord en quoi consiste un processus d'interconnexion de données. Ensuite, nous présentons les techniques les plus communément utilisées dans la littérature, à savoir l'alignement d'ontologies (*ontology alignment*) et la liaison de données (*data interlinking*). Enfin, nous listons les différentes stratégies d'interconnexion de données.

Ce chapitre est donc organisé comme suit. Les sections 2.2, 2.3 et 2.4 détaillent

respectivement ces trois axes de recherche : la RI dans les environnements connectés, la RI dans les corpus documentaires et l'interconnexion de données hétérogènes. Nous concluons ce chapitre par un bilan ainsi qu'une discussion dans la section 2.5.

2.2 La recherche d'information dans les environnements connectés

Bien qu'ils aient été largement explorés dans la littérature [33, 4, 143, 143, 179], peu de travaux proposent une définition formelle des environnements connectés. Nous pouvons en citer trois [3, 115, 119]. Ejaz et al. [3], définissent un environnement connecté comme *un petit monde interconnecté où les dispositifs disposant de capteurs travaillent en collaboration pour rendre la vie des humains plus confortable*. McGlenn et al. [119], proposent une définition similaire, ils suggèrent *qu'un environnement connecté est une infrastructure capable d'acquérir et d'appliquer des connaissances sur l'environnement et ses habitants afin d'améliorer leur expérience de vie*. Mansour et al. [115], proposent la définition formelle suivante :

Définition 1. *Un environnement connecté est une infrastructure qui héberge des réseaux de capteurs capables de fournir des données importantes utilisables par diverses applications.*

$$EC = \bigcup_{i=0}^n c_i \quad \forall i \in \mathbb{N}$$

Où :

- *EC est l'acronyme d'environnement connecté et c_i est une instance de capteurs appartenant à EC.*

Nous adoptons cette définition dans la suite de ce chapitre. Les instances de capteurs c_i appartenant à l'environnement connecté peuvent être statiques ou mobiles (ex. un capteur installé sur un mur, un capteur installé dans un téléphone mobile, etc.), accompagnés ou pas de leurs métadonnées (ex. la portée du capteur, sa localisation, la précision des valeurs mesurées, etc.), mesurant des données multimédias (image, son ou vidéo) ou scalaires (ex. un nombre entier, un nombre réel, etc.). Pour gérer tous ces aspects et permettre aux utilisateurs des EC d'exploiter facilement les données de capteurs, divers types de systèmes ont été proposés, par exemple, les systèmes de gestion de bases de données relationnelles, les systèmes de gestion de bases de données NoSQL et aussi d'autres systèmes plus complexes qui modélisent la sémantique. Dans ce qui suit, nous présentons ces systèmes sous l'angle de la RI classique et de la RI sémantique dédiées aux environnements connectés respectivement dans les sections 2.2.1 et 2.2.2. Nous concluons cette première partie par un bilan et une discussion dans la section 2.2.3.

2.2.1 La recherche d'information classique dans les environnements connectés

Une méthode classique de la gestion et du requêtage des données de capteurs consiste à stocker ces données dans des **systèmes de gestion de bases de données relationnelles** (SGBDR) [142, 171]. Les SGBDR sont utilisés pour stocker et récupérer d'immenses quantités de données. Ils ont un schéma prédéfini et stockent les données dans des lignes et des colonnes de tables. Toutefois, bien que les langages de requêtage pour les SGBDR couvrent un vaste éventail de types de requêtes (requêtes SQL de sélection, agrégation, jointure, etc.), la nature dynamique des données de capteurs (observations produites dynamiquement dans **un flux de données temporel**) nécessite de nouveaux types de requêtes non prises en charge par les SGBDR.

Un flux de données temporel est défini comme une séquence non limitée de valeurs, chacune portant un horodatage qui indique quand l'observation a été produite [24]. On parle également dans d'autres travaux de **série temporelle** (série de données indexée par le temps) [132, 184, 199]. Si on prend l'exemple d'un capteur de lumière *LightSensor01*, le flux de données généré par ce dernier est une suite de tuples <heure, niveauDeLumière> (ex. <12:00:00,30>, <12:00:10,32>,...). Alors que les requêtes du type SQL conviennent aux données archivées dans des bases de données, les flux de données nécessitent des requêtes continues portant sur de longues durées qui traitent et produisent des résultats au fur et à mesure de l'arrivée des observations. Par exemple, l'utilisateur pourrait avoir besoin de faire la requête suivante : je souhaite avoir la valeur maximale de la température durant les 15 dernières minutes dans le deuxième étage du bâtiment. Ce type de requête qui nécessite (i) une gestion du flux de données en temps réel, ainsi qu'une (ii) description détaillée de l'aspect spatio-temporel, n'est pas géré par les systèmes de gestion de bases de données classiques [164, 135].

Pour résoudre ces problèmes, plusieurs travaux ont été conduits pour proposer des extensions du langage SQL. Dans [90, 164, 9], les auteurs proposent trois extensions qui introduisent toutes la notion de **fenêtre**. Une fenêtre fait référence à la durée de temps la plus récente fixée par l'utilisateur, par exemple, les 15 dernières minutes. Ces travaux consistent essentiellement à proposer de nouveaux opérateurs et de nouveaux types de données pour gérer l'aspect dynamique des requêtes. De la même façon, les auteurs dans [102, 123, 60, 100] proposent quatre langages de requêtage inspirés du langage SQL pour gérer les aspects spatio-temporel des requêtes sur les données de capteurs. Pour ce faire, ils introduisent également de nouveaux opérateurs.

Outre le fait que les SGBDR ne sont pas adaptés à la gestion de flux de données, ils présentent une autre limite, celle de leur capacité de passage à l'échelle (ou scalabilité) limitée [175, 53]. En effet, les capteurs dans les EC produisent généralement des mesures à des intervalles réguliers (par exemple, chaque minute), la quantité de données à traiter devient rapidement très grande et un seul serveur n'est plus suffisant ni en terme de stockage, ni en terme de puissance de calcul pour gérer toutes ces données. La solution qui semble la plus appropriée est d'ajouter plus de serveurs pour y stocker les nouvelles données. Cependant, la gestion des bases de données relationnelles distribuées est très

complexe et difficile à mettre en oeuvre [175].

Pour résoudre ce problème, **les systèmes de gestion de bases de données NoSQL (SGBDNoSQL)** ont vu le jour [175, 53]. Les SGBDNoSQL répartissent généralement leurs données sur plusieurs serveurs. En cas d'augmentation du volume de données, de nouveaux serveurs peuvent alors facilement être ajoutés. Ils peuvent ainsi enregistrer et traiter sans problème de gros volumes de données de capteurs. Les SGBDNoSQL ne stockent pas les données selon un schéma fixe comme c'est le cas pour les SGBDR. Par exemple, il pourrait s'agir d'un schéma de stockage clé-valeur, où chaque donnée enregistrée a une clé pour l'identifier et une valeur qui peut être n'importe quel type d'objet (ex. nombre, chaîne de caractères, tableau, etc.). Les lecteurs qui souhaiteraient avoir de plus amples informations peuvent consulter les travaux [120, 170]. Étant donné que les SGBDNoSQL ne stockent pas les données suivant un schéma fixe, les langages d'interrogation proposés sont moins puissants qu'SQL [175] et donc des problèmes de requêtage sont également rencontrés ici par les utilisateurs. Par conséquent, un nouveau type de systèmes de gestion de bases de données dédié spécifiquement aux flux de données temporelles est apparu dans la littérature, à savoir, **les systèmes de gestion de bases de données de séries temporelles (SGBDST)** [83].

Les systèmes de gestion de bases de données de séries temporelles sont spécialisés dans le stockage et l'interrogation des données de séries temporelles, que ce soit des données de capteurs ou d'autres types de données (ex. données financières, transactions boursières) [130]. La majorité des SGBDST se basent sur des SGBNoSQL déjà existants pour stocker les séries temporelles ou bien ils adoptent eux-mêmes le paradigme NoSQL [16] et donc ils répondent au besoin de passage à l'échelle. Par exemple, les trois SGBDST Blueflood¹, KairosDB² et NewTS³ utilisent le SGBDNoSQL Cassandra pour stocker les séries temporelles, tandis que, InfluxDB⁴, Kdb+⁵ et Prometheus⁶ proposent leurs propres systèmes de stockage basés sur le paradigme NoSQL. Les SGBDST sont optimisés pour gérer et stocker les données temporelles avec précision. Cela se traduit par de meilleures performances en matière de stockage de données, mais aussi d'interrogation [131]. De plus, les SGBDST proposent leurs propres langages de requêtage. Ces langages permettent de faire plusieurs types de requête dédiés aux séries temporelles, telles que les requêtes continues qui sont recalculées au fur et à mesure que de nouvelles données sont ajoutées et les requêtes d'agrégation qui permettent d'avoir le résumé de l'évolution d'une valeur sur une large période en un temps de traitement de quelques millisecondes (par exemple, l'évolution de la température ce mois-ci par rapport à la même période des six derniers mois) [131].

Enfin, il existe de nombreux systèmes et plateformes industriels open-source qui s'appuient sur les systèmes de gestion de bases de données susmentionnés (SGBDR,

1. <http://blueflood.io/>
2. <https://kairosdb.github.io/>
3. <http://opennms.github.io/newts/>
4. <https://www.influxdata.com/>
5. <https://code.kx.com/>
6. <https://prometheus.io/>

SGBDNoSQL et SGBDST) et qui sont dédiés spécifiquement à la gestion de réseaux de capteurs [167, 84, 39]. Nous pouvons citer par exemple la plateforme ThingsBoard⁷. ThingsBoard est une plateforme open-source dédiée à la collecte, au traitement et à la visualisation des données de capteurs. La plateforme donne la possibilité à l'utilisateur de choisir le mode de stockage qui lui convient, soit le SGBDR PostgreSQL⁸, le SGBD-NoSQL Cassandra⁹ ou le SGBDST Timescale¹⁰. La performance des opérations dépend alors de la taille du réseau de capteurs et du mode de stockage choisi. Thingsboard propose également une API REST qui permet d'interroger facilement les données de capteurs quel que soit le mode de stockage choisi. L'API permet de faire plusieurs types de requêtes, telles que par exemple les requêtes continues, les requêtes d'agrégation et les requêtes sur des fenêtres temporelles.

Pour résumer, dans cette section nous avons passé en revue les méthodes classiques de la gestion des données de capteurs, à savoir, l'utilisation des SGBDR, des SGBDNoSQL et des SGBDST. Chacune de ces méthodes a ses avantages et ses limites, néanmoins, les SGBDST restent les plus adaptés aux données de capteurs vu qu'ils ont été spécifiquement conçus pour ce type de données.

Dans la section suivante, nous étudions les approches sémantiques associées à la recherche d'information pour les réseaux de capteurs et ce qu'elles amènent de plus par rapport aux approches classiques que nous venons de voir.

2.2.2 La recherche d'information sémantique dans les environnements connectés

Dans les environnements connectés, les capteurs installés proviennent généralement de plusieurs fournisseurs. Les observations produites ont des formats, vocabulaires, précisions et unités de mesures hétérogènes, ce qui rend difficile le partage et l'exploitation en commun de ces données. Chaque capteur peut avoir sa propre façon de nommer des informations sémantiquement identiques, par exemple, "température" et "température moyenne", ou bien "CO2" et "dioxyde de carbone". Si l'utilisateur souhaite établir une requête pour obtenir les niveaux de températures dans un environnement où un réseau de capteurs hétérogènes est déployé, ce dernier doit connaître en détail les noms des propriétés mesurées par chaque capteur. Par conséquent, la recherche sur les données de capteurs devient complexe. La figure 2.1 montre deux capteurs (Light-SensorX48 et LM393) qui ont des schémas différents, bien qu'ils mesurent tous les deux les niveaux de lumière.

Afin de maîtriser l'aspect hétérogène des réseaux de capteurs et permettre aux utilisateurs de les interroger facilement, un ensemble de travaux de recherche a été mené pour proposer des approches de recherche d'information sémantique (RIS) pour les réseaux de capteurs [19, 24, 187, 181, 139]. Ces approches consistent à combiner les

7. <https://thingsboard.io/>

8. <https://www.postgresql.org/>

9. <https://cassandra.apache.org/>

10. <https://www.timescale.com/>

```

LightSensorX48: {light_level FLOAT, timestamp TIMESTAMP}
LM393: {luminosity Double, timed DATETIME}
    
```

FIGURE 2.1 – Schémas de flux de capteurs hétérogènes

données de capteurs avec les techniques du Web sémantique. Ce mariage est communément désigné dans la littérature sous le nom de *Web sémantique de capteurs* (semantic sensor web) [159].

Contrairement à la RI sur les corpus documentaires, la RI sur le Web sémantique de capteurs est un domaine de recherche récent, il n'existe pas vraiment de travaux de référence que l'on pourrait citer et que l'on pourrait exploiter pour présenter le processus classique de la recherche sémantique pour les réseaux de capteurs. Nous nous sommes donc basés sur les travaux [19, 24, 187, 181, 139] pour proposer une synthèse des étapes importantes du processus de la RIS pour les réseaux de capteurs. Cette synthèse est présentée dans la figure 2.2.

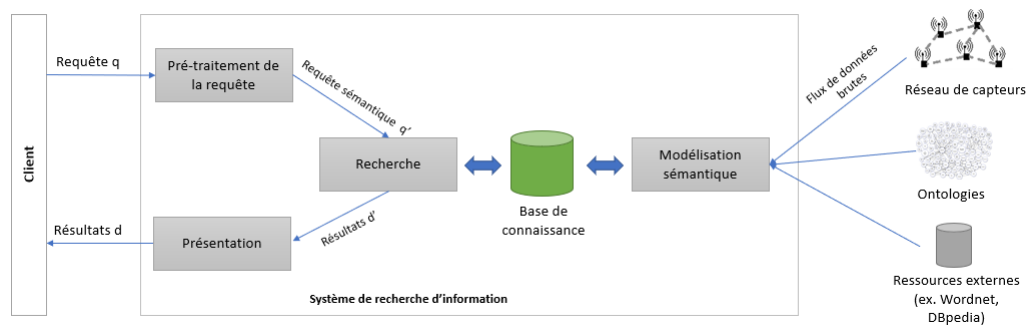


FIGURE 2.2 – Le processus de la RIS pour les réseaux de capteurs

Généralement, la RIS pour les réseaux de capteurs repose sur les étapes principales listées ci-dessous :

- **Modélisation sémantique** : cette étape consiste à traduire le flux hétérogène des données de capteurs en une représentation sémantique (ex. triplets RDF [24]). Pour ce faire, plusieurs ontologies et ressources externes peuvent être utilisées [19]. Ensuite, ces données sont stockées et gérées par des systèmes de gestion de bases de données similaires à ceux présentés dans la section 2.2.1 (ex. SGBDR [181]).
- **Traitement de la requête** : la requête de l'utilisateur est souvent exprimée sous forme d'une liste de mots-clés ou une phrase en langage naturel. Parfois, elle est trop courte ou bien mal formulée. Cette étape a pour but d'analyser la requête afin de mieux comprendre et cibler le besoin de l'utilisateur, puis la traduire en un format compréhensible par le système (ex. requête SPARQL [24], requête SQL [181]).

- **Recherche** : une fois la requête pré-traitée, celle-ci est exécutée sur la base de connaissance et les résultats sont retournés.
- **Présentation** : le système présente les résultats à l'utilisateur en réponse à sa requête initiale dans un format compréhensible (par exemple, un tableau avec l'ensemble des observations sélectionnées).

Dans cette thèse, nous nous intéressons principalement à la modélisation sémantique des données de capteurs (hypothèse 1, section 1.2.3). Par conséquent, dans ce qui suit, nous détaillons uniquement cette étape.

La tâche de modélisation des données de capteurs (les observations produites) et leurs métadonnées (ex. nom du capteur, nom de la propriété, unité de mesure, etc.) avec des ontologies a été beaucoup abordée par la communauté du Web sémantique [32, 18, 31, 134, 8, 153]. Toutefois, la majorité des premières approches se sont concentrées uniquement sur la modélisation des métadonnées des capteurs, sans tenir compte de la description des observations [32]. De plus, une grande partie de ces approches sont souvent spécifiques à un projet de recherche scientifique, parfois, l'ontologie n'est pas disponible en téléchargement ou bien carrément abandonnée après la fin du projet en question [99, 78].

Pour résoudre ces problèmes et proposer des ontologies génériques permettant de modéliser n'importe quel type de réseau de capteurs, plusieurs travaux ont été conduits [31, 153, 8, 116]. Nous présentons dans ce qui suit, quelques-unes de ces ontologies. Tout d'abord, le W3C (World Wide Web Consortium)¹¹ a proposé un nouveau standard : l'ontologie **SSN** (Semantic Sensor Network) [31]. SSN est une ontologie générique et indépendante du domaine, elle permet à la fois la modélisation des métadonnées (par exemple, les concepts *Property* et *BatteryLifetime*) et des observations (par exemple, les concepts *Observation* et *Result*). Néanmoins, l'ontologie SSN présente plusieurs lacunes : (i) elle ne permet pas de décrire précisément l'environnement où les capteurs sont déployés (ex. salles, étages, bâtiment, etc.); (ii) elle ne contient pas non plus les concepts permettant de modéliser l'aspect de la mobilité des capteurs (un capteur peut par exemple être installé sur un drone ou sur un téléphone mobile); (iii) elle ne permet pas de modéliser les données et les propriétés multimédias (image, son ou vidéo); (iv) il est difficile de décrire certaines métadonnées de capteurs, telles que la portée et la précision, vu qu'il y a une seule propriété *Description* pour regrouper toutes ces informations [18]; (v) l'ontologie SSN ne permet pas de modéliser les événements pouvant être détectés par les capteurs installés dans l'environnement. Dans [153], les auteurs proposent l'ontologie **OntoSensor** qui étend l'ontologie SUMO (The Suggested Upper Merged Ontology) [134] et utilise des termes issus des deux normes ISO 19115¹² et SensorML¹³. Cette ontologie est générique, elle propose une description très détaillée des métadonnées des capteurs : les différentes catégories de capteurs (ex. capteur électromagnétique et capteur chimique), leur configuration (ex. fréquence

11. <https://www.w3.org/>

12. <http://bit.ly/2nedvM1>

13. <http://www.opengeospatial.org/standards/sensorml>

de mesure des données) et leur capacité (ex. portée, précision, échange de données). Cependant, l'ontologie est volumineuse, d'une grande complexité et ne contient pas de concepts décrivant les observations ni les événements. Dans [8], les auteurs proposent l'ontologie **MSSN-onto** (Multimedia SSN Ontology) qui étend l'ontologie SSN avec les concepts et les relations concernant les aspects techniques des données multimédias (par exemple, les concepts *VideoSensor*, *AudioSensor* et *ImageSensor*, et les propriétés *VideoData*, *AudioData* et *ImageData*) ainsi que la définition d'événements (par exemple, les concept *Event* et *EventStatement*). Cependant, les propriétés, les observations scalaires ou multimédias produites, et leurs métadonnées respectives manquent de clarté. De plus, les auteurs font la supposition que les capteurs ne changent pas de position. L'ontologie ne contient donc pas les concepts permettant de décrire la mobilité des capteurs. MSSN-Onto ne permet pas non plus de décrire les types d'infrastructures pouvant héberger les capteurs (ex. bâtiment, téléphone portable, drone, etc.). SSN/SOSA. Les auteurs dans [71] proposent **SOSA/SSN**¹⁴ un ensemble d'ontologies publiées à la fois comme recommandation du W3C (World Wide Web Consortium) et comme norme de mise en œuvre de l'OGC (Open Geospatial Consortium). Cet ensemble d'ontologies comprend un module de base léger appelé SOSA (Sensor, Observation, Sampler, et Actuator) et un module d'extension plus expressif appelé SSN (Semantic Sensor Network). Ensemble, ils définissent les différentes composantes des environnements connectés (capteurs, actionneurs, observations, propriétés mesurées, etc.). Dans SOSA/SSN, les capteurs sont déployés sur une plate-forme via un certain processus de déploiement. Les plates-formes peuvent être des infrastructures du monde réel (par exemple, un bâtiment) ou des dispositifs électroniques (par exemple, des téléphones mobiles). Cependant, la représentation des différents types de plateformes n'est pas détaillée. Enfin, SOSA/SSN propose une représentation simple des nœuds de capteurs, ainsi que des systèmes/dispositifs (de détection). Cependant, les auteurs ne proposent aucun concept lié à la mobilité pour représenter les capteurs mobiles, ni les données/propriétés multimédia. Dans [116], les auteurs proposent l'ontologie **HSSN** qui étend l'ontologie SSN. HSSN introduit un ensemble de concepts et de relations permettant de décrire la diversité des infrastructures, des capteurs et des données. Nous détaillons ci-dessous chacun de ces points.

- **La diversité des infrastructures** : l'ontologie SSN ne propose qu'un seul concept *Platform* pour représenter les entités qui peuvent héberger des capteurs, elle ne permet donc pas de faire la distinction entre les plateformes classiques (ex. bâtiment, bureau, mur, etc.) et les appareils incluant des capteurs intégrés (ex. téléphone portable, machine à laver, drone, etc.). L'ontologie HSSN étend cette représentation pour prendre en compte la diversité des plateformes. Elle introduit les concepts *Infrastructure* et *Device* pour représenter respectivement les environnements physiques et les appareils électroniques pouvant héberger des capteurs. De plus, HSSN introduit les concepts et les relations permettant de décrire précisément les environnements physiques, leurs différentes sous-localisations et

14. <https://www.w3.org/TR/vocab-ssn/>

les relations spatiales les reliant. En effet, chaque infrastructure (par exemple, un bâtiment connecté) est décrite par une carte de localisation représentée par le concept *LocationMap*. Le concept *LocationMap* contient la propriété *isComposeOf* qui permet d'inclure plusieurs sous-localisations (par exemple, des étages et des bureaux). Les liaisons spatiales entre ces sous-localisations (par exemple, au-dessus, inclut-dans) sont représentées par plusieurs propriétés, telles que par exemple *isAbove* ou *contains*.

- **La diversité des capteurs** : HSSN introduit principalement les deux concepts *MobileSensor* et *StaticSensor* pour classer les capteurs dans deux grandes catégories (capteurs statiques et capteurs mobiles). Afin de suivre le déplacement des capteurs mobiles dans l'environnement, HSSN introduit également un ensemble de propriétés telles que par exemple *isCurrentlyLocatedAt* et *hasPastLocation*. De plus, HSSN propose d'ajouter plusieurs concepts pour décrire la zone couverte par chaque capteur (ex. le concept *CoverageArea*). Néanmoins, plusieurs autres métadonnées, telles que par exemple la précision des valeurs mesurées par les capteurs ou leur capacité à communiquer et échanger des données ne sont pas modélisées dans HSSN.
- **La diversité des données** : HSSN intègre les concepts et propriétés relatifs aux données multimédias (audio, image et vidéo) qui peuvent être mesurées par des capteurs mobiles ou statiques. Pour ce faire, les auteurs se basent sur l'ontologie MSSN-onto citée plus haut qui décrit les aspects techniques/métadonnées des objets multimédias. HSSN utilise plusieurs concepts et propriétés issus de l'ontologie MSSN-onto et complète ceux-ci par des concepts et des propriétés supplémentaires tels que par exemple, les propriétés *mediaSenses* et *scalarSenses*.

En résumé, les approches de la RIS pour les réseaux de capteurs s'appuient sur les techniques du Web sémantique pour faire face à l'hétérogénéité des données de capteurs et faciliter l'interrogation de ces données par les utilisateurs. Ces approches sont généralement basées sur un processus qui consiste en quatre étapes : modélisation sémantique, traitement des requêtes, recherche et présentation des résultats. L'étape de la modélisation sémantique est celle à laquelle nous nous intéressons dans cette thèse. Elle se base généralement sur des ontologies pour représenter la sémantique dans les réseaux de capteurs. Nous avons examiné plusieurs ontologies [31, 153, 8, 116], l'ontologie HSSN [116] semble être la plus aboutie même si elle présente certaines lacunes auxquelles nous ferons face en l'étendant (cf. section 2.2.3).

Après avoir passé en revue les approches de la RI classique (section 2.2.1) et de la RI sémantique (section 2.2.2) pour les réseaux de capteurs, dans la section suivante, nous faisons une brève synthèse de ce que nous avons vu dans ces deux sections. Ensuite, nous présentons à travers une discussion ce que nous retenons dans le cadre de cette thèse.

2.2.3 Synthèse et discussion

Dans cette première partie du chapitre état de l'art, nous avons passé en revue les travaux autour de la RI classique (section 2.2.1) et de la RI sémantique (section 2.2.2) pour les réseaux de capteurs.

Tout d'abord, concernant la RI classique pour les réseaux de capteurs (section 2.2.1), nous avons vu qu'il y a trois tendances principales, à savoir, l'utilisation des SGBDR, des SGBDNoSQL et des SGBDST. Les SGBDR souffrent de problèmes de passage à l'échelle très limité et de langage de requêtage (principalement SQL) non adapté aux données de capteurs. Les SGBDNoSQL résolvent le problème de passage à l'échelle, dans le cas de données très volumineuses, ils sont beaucoup plus performants. Toutefois, les langages de requêtage dédiés aux SGBDNoSQL sont moins puissants que ceux des SGBDR et toujours non adaptés aux données de capteurs. Les SGBDST sont spécifiquement conçus pour le stockage et l'interrogation des données de séries temporelles (par exemple, les données de capteurs), ils répondent au besoin de passage à l'échelle (ils se basent généralement sur le paradigme NoSQL) et proposent des langages de requêtage dédiés aux séries temporelles.

La RIS dédiée aux réseaux de capteurs consiste à combiner les données de capteurs avec les techniques du Web sémantique (ex. les ontologies). Elle repose sur un processus classique qui commence par une étape de modélisation sémantique à laquelle nous nous intéressons dans cette thèse. La modélisation sémantique consiste à traduire le flux hétérogène des données de capteurs en une représentation sémantique qui s'appuie sur des ontologies [31, 153, 8, 116]. Le tableau 2.1 récapitule l'ensemble des ontologies présentées et les évalue en se basant sur les sept critères suivants : la généricité (indépendance du domaine d'application, colonne 2), la modélisation des métadonnées des capteurs (colonne 3), la modélisation des observations (colonne 4), la modélisation des infrastructures pouvant héberger des capteurs (colonne 5), la modélisation de la mobilité des capteurs (colonne 6), la modélisation des données multimédias (colonne 7) et la modélisation des événements (colonne 8). Nous avons choisi ces critères parce que nous pensons qu'ils sont les plus importants dans notre contexte.

Nous utilisons les symboles suivants pour évaluer les ontologies : "✓" pour exprimer une couverture complète, "✗" pour exprimer une absence de couverture et "Partiel" pour exprimer une couverture partielle.

Ontologies	Généricité	Méta-données	Observations	Infra-structure	Mobilité des capteurs	Données multimédias	Événements
SSN [31]	✓	Partiel	✓	Partiel	✗	✗	✗
OntoSensor [153]	✓	✓	✗	✗	✗	✗	✗
MSSN-onto [8]	✓	Partiel	✓	✗	✗	✓	Partiel
HSSN [116]	✓	Partiel	✓	✓	✓	✓	✗

TABLE 2.1 – Évaluation des ontologies réseau de capteurs

Comme vous pouvez le constater dans le tableau 2.1, aucune des quatre ontologies ne couvre complètement les sept critères d'évaluation. Néanmoins, l'ontologie HSSN

est celle qui couvrent le plus de critères, nous choisissons donc cette ontologie pour modéliser les données du réseau de capteurs. Pour compléter l'ontologie HSSN, nous proposons de l'étendre, ce processus sera détaillé dans le chapitre suivant.

Enfin, dans notre contexte, celui de l'explication des événements dans les environnements hybrides, les réseaux de capteurs sont une source importante de données qui doit être complétée par des données supplémentaires issues du corpus de documents. Dans ce cadre, une grande limite reste non résolue : les systèmes présentés dans cette première partie sont spécifiques aux données de capteurs et ne permettent pas de gérer à la fois les données de capteurs et les corpus documentaires, en d'autres mots, il y a une sorte de fossé entre ces deux mondes. Nous essaierons de résoudre cette problématique en exploitant les représentations sémantiques (hypothèse 1., section 1.2.3), les données issues des événements et l'ontologie HSSN étendue.

Maintenant que nous avons exploré les travaux concernant la recherche d'information dans les environnements connectés, nous passons dans la section suivante à la recherche d'information sur les corpus documentaires.

2.3 La recherche d'information dans les corpus documentaires

Un corpus est défini dans le dictionnaire Larousse comme "*un recueil de documents relatifs à une discipline, réunis en vue de leur conservation*"¹⁵. Cependant, dans les SI dédiés aux grands environnements tels que par exemple, les centres commerciaux ou les grandes entreprises, les corpus documentaires sont de plus en plus de nature hétérogène incluant des documents relatifs à plusieurs disciplines (par exemple, gestion des ressources humaines, gestion des fichiers des abonnés, gestion des équipements, etc.). Nous proposons donc la définition suivante :

Définition 2. *Un corpus est un ensemble de documents (textes, images, vidéos, etc.) regroupés dans un but précis d'analyse.*

$$CD = \bigcup_{i=0}^n d_i \forall i \in \mathbb{N}$$

Où :

- **CD** est l'acronyme de corpus documentaire et d_i est un document appartenant à CD.

Bien que les systèmes de la RI (classique et sémantique) pour les corpus documentaires ait considérablement évolués avec l'émergence des moteurs de recherche [23] et des techniques du Big Data [80], ils reposent principalement sur un processus classique. Ce processus est constitué de cinq étapes principales (Figure 2.3) :

15. <https://www.larousse.fr/dictionnaires/francais/corpus/19410>

16. Inspiré du processus général de la RI présenté dans [17]

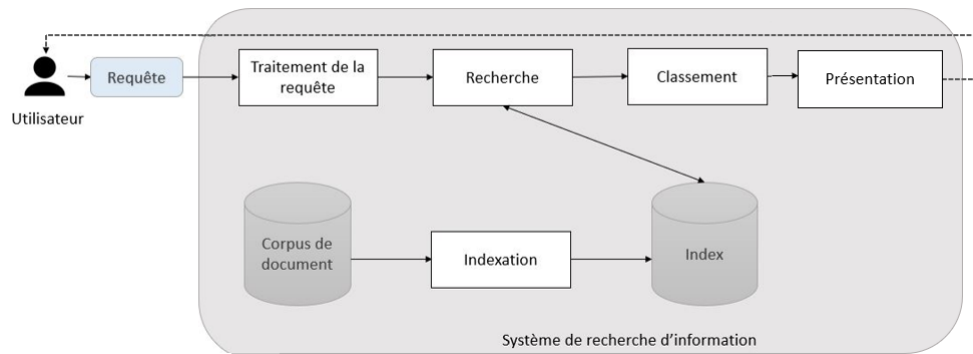


FIGURE 2.3 – Le processus classique des systèmes de la RI pour les corpus documentaires¹⁶

- **Indexation** : avant que l'utilisateur puisse interroger le système, l'indexation de l'ensemble des documents est nécessaire. Cette première étape a pour but d'analyser le corpus et de générer pour chaque document une représentation logique réduite (index) [65]. Par exemple, une structure d'indexation utilisée par de nombreux systèmes de la RI est *l'index inversé*. Cette méthode représente les documents sous forme d'une liste de termes et de poids (le nombre de fois où le terme apparaît dans le document).
- **Traitement de la requête** : cette étape a pour but d'analyser la requête afin de mieux comprendre et cibler le besoin de l'utilisateur, puis la transformer en un format logique compréhensible par le système (similaire à celui utilisé pour l'indexation des documents [17]).
- **Recherche** : cette étape procède à l'appariement entre la représentation de la requête et le contenu de l'index. Cet appariement permet de retrouver, via l'index les documents jugés pertinents par rapport à la requête. Le système récupère ainsi les documents ou les sous-parties des documents qui répondent au besoin de l'utilisateur. Pour ce faire, plusieurs modèles de recherche peuvent être adoptés (ex. le modèle vectoriel, le modèle probabiliste).
- **Classement** : cette étape a pour but de classer les documents retournés par le module de recherche par ordre de pertinence.
- **Présentation** : le système présente les résultats à l'utilisateur en réponse à sa requête initiale, dans une interface graphique et dans un format compréhensible.

Dans ce qui suit, nous présentons les systèmes de la RI classique et de la RI sémantique dédiés aux corpus documentaires respectivement dans les sections 2.3.1 et 2.3.2. La section 2.3.3, fait le point sur les systèmes question-réponse. Nous mettons l'accent sur les approches 5W1H dans la section 2.3.4 étant donné que nous l'utilisons dans notre proposition (hypothèse 2 et 3, section 1.2.3). La section 2.3.5 passe en revue les métriques d'évaluation des systèmes de la recherche d'information. Nous concluons par une discussion dans la section 2.3.6.

2.3.1 La recherche d'information classique dans les corpus documentaires

Plusieurs modèles de la RI pour les corpus documentaires ont été proposés dans la littérature. Les plus évoqués sont ceux présentés par Singhal dans son article [160], à savoir les modèles booléen, vectoriel et probabiliste. **Le modèle booléen** est basé sur l'algèbre booléenne et l'utilisation des opérateurs AND, OR et NOT [155]. Les requêtes sont formulées en termes d'expressions booléennes, par exemple, pour la requête suivante : "jus ET orange (PAS pomme)", le modèle booléen retourne les documents contenant les termes "jus" et "orange", mais pas le terme "pomme". **Le modèle vectoriel** permet d'attribuer des poids non binaires aux termes de l'index dans les requêtes et les documents [155]. Dans ce modèle, un document textuel est représenté par un vecteur de fréquences de termes apparaissant dans ce document. Afin de mesurer la similarité entre deux documents, on calcule la similarité entre leurs vecteurs de fréquence respective. Enfin, **le modèle probabiliste** vise à classer les documents par probabilité de pertinence par rapport à la requête de l'utilisateur [155]. La probabilité de pertinence est calculée en fonction de différentes caractéristiques des documents (par exemple, le sac de mots, la longueur du document).

En plus de ces trois modèles qui sont largement utilisés dans la littérature et qui ont chacun leurs avantages et leurs limites, plusieurs travaux proposant des extensions ou des combinaisons de ces modèles ont été réalisés. Par exemple, Salton et al. [155] étaient parmi les premiers à proposer d'étendre le modèle booléen. Pour ce faire, ils combinent les caractéristiques du modèle vectoriel avec les propriétés algébriques du modèle booléen de manière à couvrir la correspondance partielle des documents avec les requêtes. Turtle et Croft [172] sont également les premiers à introduire l'utilisation des réseaux bayésiens dans la RI pour représenter les dépendances probabilistes entre un document et une requête sous la forme d'un graphe de dépendances. Ils combinent les modèles de recherche booléens et probabilistes associés avec des méthodes statistiques. Les auteurs dans [34, 59] proposent deux approches similaires qui consistent à combiner efficacement les résultats de plusieurs systèmes de recherche ayant des modèles différents. Le système donne des poids d'importance aux différents modèles de recherche en fonction des retours de l'utilisateur. Au fur et à mesure que l'utilisateur effectue des retours, ces poids sont modifiés pour améliorer les futurs résultats. Ces deux travaux constituent les premières tentatives dans la littérature de l'utilisation des algorithmes d'apprentissage automatique (AA) dans les systèmes de recherche.

Les algorithmes d'AA utilisent (i) un ensemble de *caractéristiques* (features) décrivant le corpus documentaire ainsi que les requêtes et (ii) un *ensemble d'entraînement* constitué d'une liste de requêtes et des documents pertinents que le système doit retourner pour chacune de ces requêtes. Généralement, les caractéristiques peuvent être extraites directement du texte (ex. fréquence des mots, longueur du texte) ou en utilisant des techniques du traitement automatique des langues NLP (ex. racines des mots, étiquetage grammatical). En se basant sur toutes ces informations, le système apprend à classer les documents comme pertinents ou non pertinents pour les nouvelles requêtes

[114]. Plusieurs méthodes d'AA ont été développées et appliquées à la RI sur les corpus documentaires, telles que *SVM* (Support Vector Machines) [94, 203, 25, 11] ou le *réseau de neurones* [121, 103]. Les évaluations de ces systèmes démontrent une amélioration des performances comparées aux systèmes classiques (booléen, vectoriel et probabiliste) [114]. Toutefois, cette amélioration est étroitement liée à la taille de l'ensemble d'entraînement. Plus l'ensemble d'entraînement est grand et plus il est diversifié (documents portant sur différents contextes), meilleure est la performance du système. De plus, l'application des AA dans la RI peut être complexe et coûteuse en termes de ressources, notamment pour l'étape de l'entraînement du modèle.

Pour résumer, dans cette section nous avons passé en revue différents modèles de la RI classique dans les corpus documentaires, à savoir, les modèles booléen, vectoriel, probabiliste, les modèles qui étendent ou combinent ces trois modèles et les modèles basés sur des algorithmes d'apprentissage automatique. Les modèles basés sur des algorithmes AA sont ceux qui ont la meilleure performance [114]. Dans la section suivante, nous verrons les approches sémantiques de la RI pour les corpus documentaires et en quoi elles apportent des améliorations par rapport aux approches classiques.

2.3.2 La recherche d'information sémantique dans les corpus documentaires

La digitalisation des entreprises 1.1.2 ainsi que l'évolution du réseau Internet [82] sont parmi les phénomènes qui se développent le plus rapidement dans l'ère de l'information impliquant une énorme quantité de données. Dans ce contexte, les systèmes de la RI classique (section 2.3.1) ont pour but d'accompagner les utilisateurs dans leurs recherches. Cependant, ces systèmes sont généralement loin de satisfaire complètement les besoins d'information exprimés [148]. A titre d'exemple, supposons qu'un utilisateur souhaite trouver des informations sur "Lincoln", la célèbre marque de voiture, en insérant dans un moteur de recherche le mot-clé "Lincoln", il aura des pages liées au domaine de l'automobile et des pages sur l'histoire et la politique domaines. Les résultats sont fréquemment erronés ou imprécis parce qu'ils ne sont pas liés au contexte de la requête de l'utilisateur. Pour faire face à cette limite, les approches de la recherche d'information sémantique (RIS) ont émergé dans la littérature.

Certaines approches de la RIS adoptent complètement les techniques du Web sémantique [207, 88, 61], tandis que d'autres utilisent ces techniques afin d'améliorer la précision des systèmes de la RI traditionnelle [166, 173]. Le niveau de la représentation sémantique employée dans les deux approches peut varier de structures de connaissances sémantiques plutôt légères (ex. taxonomies, thésaurus, etc.) à des structures de connaissances sémantiques plus complexes (ex. réseaux sémantiques et ontologies). Selon le niveau de la représentation sémantique, trois familles d'approches principales peuvent être distinguées dans la littérature [54, 55] : les approches de l'analyse sémantique latente, les approches basées sur la conceptualisation linguistique et les approches basées sur les ontologies. Dans ce qui nous présentons brièvement ces trois approches, ensuite, nous prenons le temps de les détailler.

- **Approches de l'analyse sémantique latente (ASL)** : ces approches [85, 76] sont basées sur des techniques purement statistiques permettant de découvrir les relations de similarité entre les documents, les fragments de documents ou les mots qui apparaissent dans des collections de documents. Sur la base de ces informations, le modèle ASL regroupe les termes dans des ensembles décrivant le même *concept*. Ces approches sont celles qui traitent la sémantique de la manière la plus légère, sans aucune contribution humaine.
- **Approches basées sur la conceptualisation linguistique** : cette famille d'approches [204, 15, 147, 202, 69, 180] emploient des représentations sémantiques dites "légères", telles que les thésaurus, les taxonomies et les dictionnaires afin d'améliorer les systèmes de recherche traditionnelle.
- **Approches basées sur les ontologies** : ces approches [26, 125, 26, 40, 70, 51, 180, 108] se caractérisent par l'utilisation de représentations sémantiques très détaillées, telles que les ontologies et les réseaux sémantiques pour surmonter les limites de la RI classique.

Dans ce qui suit, nous prenons le temps de détailler chacune de ces approches. Cette partie vise à familiariser les lecteurs avec les techniques et outils dédiés à la RIS dans la littérature étant donné que nous réutilisons certains d'entre eux dans notre proposition.

— **Approches de l'analyse sémantique latente**

Dans les approches traditionnelles de la RI, les relations potentielles qui existent entre les mots-clés sont généralement ignorées. Par exemple, si la requête de l'utilisateur contient le mot *bâtiment*, et un document le mot *immeuble*, ce dernier n'est pas considéré comme pertinent par le système étant donné que la relation de synonymie n'est pas gérée par les approches traditionnelles de la RI. Le phénomène de polysémie pose également problème, c'est-à-dire que les documents qui contiennent le mot-clé de la requête mais avec une autre signification sont identifiés comme pertinents. Pour faire face à ces limites, le modèle de ASL a vu le jour [105]. Le modèle de ASL considère que les documents qui comportent un grand nombre de mots en commun sont sémantiquement proches. Pour ce faire, la co-occurrence des mots-clés dans les documents est calculée. Les mots qui apparaissent souvent ensemble sont regroupés dans des ensembles décrivant la même thématique. Lorsque le système reçoit une requête, les ensembles correspondant aux mots de la requête sont sélectionnés et utilisés ensuite pour choisir les documents pertinents. Les lecteurs intéressés par les fondements mathématiques du modèle de ASL sont invités à consulter l'article suivant [41].

— **Approches basées sur la conceptualisation linguistique**

Une autre démarche pour surmonter les limites de la RI classique sur les corpus documentaires, consiste à utiliser les ressources lexicales telles que les thésaurus, les taxonomies et les dictionnaires. Wordnet¹⁷ et Wikipedia¹⁸ sont les deux

17. <https://wordnet.princeton.edu/>

18. <https://fr.wikipedia.org/>

outils les plus communément utilisés dans la littérature [66]. Ils sont souvent utilisés pour (i) enrichir les requêtes avec plus de mots, (ii) désambigüiser le sens des termes de la requête et (iii) calculer la distance sémantique entre les mots. Par exemple, les auteurs dans [147] proposent d'utiliser WordNet pour calculer la distance sémantique entre les mots, puis d'utiliser cette distance pour calculer la similarité entre les requêtes et les documents. Dans [204, 15], les auteurs proposent deux approches pour l'enrichissement des requêtes avec les termes synonymes et hyponymes en utilisant les deux outils Wordnet et Wikipédia. L'approche proposée dans [202] utilise WordNet pour la désambigüisation du sens des termes de la requête, ensuite, elle ajoute les synonymes des mots de la requête.

— **Approches basées sur les ontologies**

L'utilisation des ontologies pour surmonter les limites de la RI classique a été considérée comme l'une des motivations principales du Web sémantique depuis son apparition à la fin des années 90 [174]. Un très grand nombre de travaux ont été proposés dans la littérature [26]. Trois approches principales peuvent être distinguées : (i) les approches interactives de formulation de requêtes, (ii) les approches d'enrichissement des requêtes et (iii) les approches d'annotation sémantique.

(i) **approches interactives de formulation de requêtes** : ces approches utilisent les ontologies comme moyen d'aider l'utilisateur dans sa recherche à travers une interface graphique. Cette interface a pour but d'aider les utilisateurs à formuler une requête précise qui répond au mieux à leurs besoins d'information, même s'ils ignorent totalement le vocabulaire du corpus documentaire [125]. Par exemple, l'interface présentée dans [26], affiche à l'utilisateur pour chaque mot de sa requête, un graphe extrait de l'ontologie de domaine qui contient le concept représentant le mot en question. L'utilisateur peut ainsi choisir à partir de ce graphe le ou les concepts qui représentent le mieux son besoin. De la même façon, les travaux présentés dans [161, 126, 200, 193, 127] proposent tous des interfaces utilisateur différentes mais avec le même principe que nous venons d'expliquer.

(ii) **approches d'enrichissement des requêtes** : ces approches utilisent les ontologies pour étendre les requêtes de l'utilisateur avec plus de mots. Par exemple, les auteurs dans [28], proposent à l'utilisateur d'étendre sa requête avec un ensemble de concepts issus des ontologies de domaine tels que par exemple, les concepts parents et les concepts fils. L'utilisateur peut ainsi formuler une requête plus spécifique ou au contraire, une requête plus générale. Les travaux présentés dans [40, 70, 51] enrichissent automatiquement la requête sans faire appel à l'utilisateur. Pour choisir les concepts, ils utilisent divers critères, tels que par exemple, la distance sémantique entre les concepts ou la concurrence des concepts dans le corpus documentaire.

(iii) **approches d'annotation sémantique** : ces approches analysent tout d'abord le corpus documentaire et construisent une ontologie ou un réseau sémantique (un ensemble de triplets RDF interconnectés) pour représenter sémantiquement

le contenu des documents. Ensuite, lorsque le système reçoit une requête, celle-ci est traduite en langage SPARQL, ou bien elle est analysée et transformée en un petit graphe RDF. Dans le cas où la requête est traduite en un graphe RDF, des techniques d'apprentissage automatique ou de comparaison de graphes sont utilisées pour sélectionner les documents pertinents. Par exemple, les auteurs dans [180, 108] proposent deux approches dédiées à la RIS. Dans ces deux travaux les documents sont traduits en un réseau sémantique en utilisant plusieurs techniques NLP et en s'appuyant sur l'outil Wordnet. Pour sélectionner les documents pertinents par rapport à la requête, les auteurs dans [180] utilisent une technique de comparaison de sous-graphes, tandis que, le système présenté dans [108] se base sur un algorithme d'apprentissage automatique. Dans [69], les auteurs présentent un système dédié à la RIS sur des ressources hétérogènes (document textuels ou documents multimédias avec des descriptions textuelles). Chaque document est traduit en une ontologie en utilisant les techniques NLP et l'encyclopédie Wikipédia. Afin de récupérer les documents pertinents, les auteurs traduisent la requête de l'utilisateur formulée en langage naturel en une requête SPARQL puis interroge les ontologies.

Pour résumer, la recherche d'information sémantique a pour but de fournir une réponse contextualisée à l'utilisateur en analysant sémantiquement son besoin (exprimé sous forme d'une requête) et le contenu du corpus documentaire. Les réponses retournées par ces systèmes sont généralement des documents accompagnés éventuellement d'extraits. Toutefois, dans notre contexte d'explication d'événements, la requête de l'utilisateur (pourquoi l'événement s'est déclenché?) nécessite des réponses contextualisées plus succinctes et plus ciblées (par exemple, les entités de l'environnement responsables du déclenchement de l'événement). Par conséquent, dans la section suivante, nous nous intéressons aux systèmes question-réponse s'appuyant sur la sémantique et qui semblent plus proches de ce que nous souhaitons proposer.

2.3.3 Systèmes question-réponse

Un système question-réponse (SQR) est réalise tâche de la RI qui consiste à fournir une réponse précise à une question ou à une déclaration formulée en langage naturel par un utilisateur [122]. Par exemple, si la question est "quelle est la date de naissance de Barack Obama?", le SQR retournera la date "4 août 1961".

Contrairement aux moteurs de recherche qui renvoient une liste de documents (éventuellement accompagnés d'extraits) à partir d'une requête saisie sous forme de liste de mots-clés, les SQR ont pour but d'extraire une réponse précise à une question dans un ensemble de documents. Ils sont caractérisés par un processus en huit étapes réparties sur trois modules (figure 2.4) :

- **Un module de traitement des questions** : ce module consiste à analyser la question de l'utilisateur et la traduire en un format mieux exploitable par un ordinateur

19. Inspiré du processus général des SQR proposé dans [5]

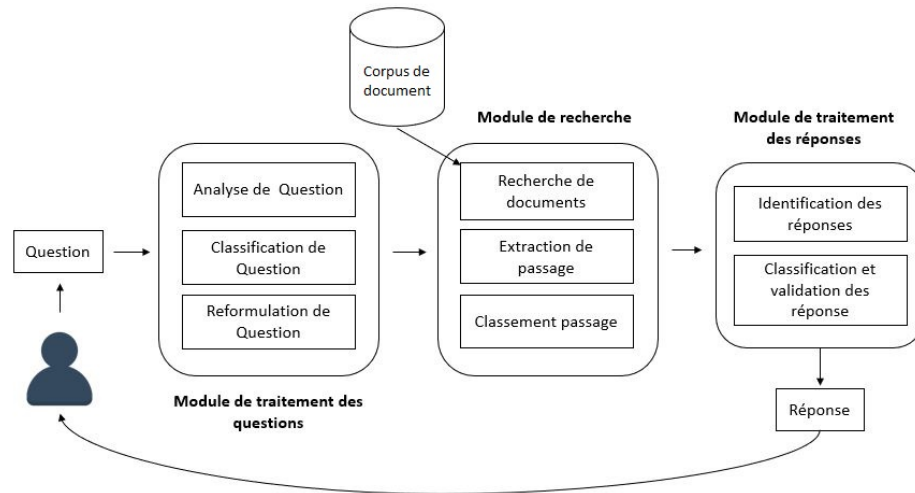


FIGURE 2.4 – Processus général des systèmes question-réponse ¹⁹

(ex, liste de mots-clés).

- **Un module de recherche** : ce dernier se base sur un système de recherche d'information (section 2.3.1 et 2.3.2) pour établir l'appariement entre la représentation de la question de l'utilisateur et celle du corpus documentaire, puis retourner tous les documents susceptibles de répondre à cette question.
- **Un module de traitement des réponses** : ce module se charge d'analyser plus finement les documents retournés par l'étape précédente, extraire la réponse, la reformuler et la retourner à l'utilisateur.

Dans ce qui suit, nous détaillons chacun de ces modules.

— **Module de traitement des questions :**

Le rôle principal du module de traitement des questions est d'appliquer tout d'abord une analyse bien détaillée de la question afin d'en extraire plusieurs informations. Selon le type d'approche, différentes techniques d'analyse de questions peuvent être appliquées [122, 43], telles que les techniques NLP (Parse tree, POS tagging, etc.) et les techniques sémantiques (identification des entités nommées, la désambiguïsation, annotation sémantique, etc.). Ces informations sont utilisées ensuite pour identifier les types des questions (par exemple, question factuelle, question de confirmation, etc.) [122], le focus de la question (un mot ou une séquence de mots qui indiquent l'élément primordial demandé dans la question, par exemple, la question "quelle est la plus haute montagne de la France?" a pour focus "la plus haute montagne") [36] et le type de réponse attendue (l'entité nommée attendue par la réponse, par exemple, lieu, date ou personne) [109]. La dernière étape du module de traitement consiste à reformuler la question sous forme d'une liste de mots-clés pertinents. Des méthodes d'enrichissement peuvent être utilisées à ce niveau pour élargir l'ensemble des mots-clés souvent en se basant sur les ressources telles que WordNet [48, 75].

— **Module de Recherche :**

Dans cette étape, la question reformulée est soumise à un système de RI (sections 2.3.1 et 2.3.2) qui va récupérer une liste classée de documents pertinents. Ensuite, afin de réduire le nombre de documents candidats et la quantité de texte à traiter, une étape de *filtrage de passage* est conduite. Le filtrage de passage repose sur le principe selon lequel les documents les plus pertinents doivent contenir les mots-clés de la question dans un nombre limité de paragraphes voisins, plutôt que dispersés sur l'ensemble du document [5]. Par conséquent, si les mots-clés sont tous trouvés dans un ensemble de N paragraphes consécutifs, cet ensemble de passages est retourné. Les SQR existants utilisent des approches très variées pour la sélection des passages. Plusieurs SQR se basent sur les méthodes de la RI qui ont été adaptées pour travailler sur des passages plutôt que sur l'ensemble des documents [133, 14, 22]. Des approches récentes introduisent les méthodes d'apprentissage supervisé [106] et par renforcement [186] pour classer les paragraphes.

Une fois la liste des passages candidats sélectionnée, une dernière étape du module de recherche consiste à attribuer un score de pertinence à ces passages en se basant sur plusieurs critères, tels que par exemple le nombre de mots de la question présents dans le passage, le nombre de mots qui séparent les mots-clés dans le paragraphe, le nombre des mots-clés manquants, ou le score du document qui contient le passage [5].

— **Module de traitement des réponses :**

Après l'extraction des passages pertinents par rapport à la question de l'utilisateur, l'étape finale consiste à extraire et valider la réponse à renvoyer. Cette étape est généralement basée sur deux sous-tâches : (i) extraction des réponses candidates : qui a pour but d'identifier et extraire les réponses des passages qui contiennent les mots-clés de la question. Pour ce faire, une approche consiste à analyser le passage à l'aide de la reconnaissance des entités nommées et l'étiquetage grammatical puis retourner le bout du texte qui correspond au type de la réponse attendue [185] (ex, lieu, date, personne) ; (ii) l'attribution de score et la validation des réponses candidates : cette étape a pour but d'identifier les réponses pertinentes à partir de la liste des réponses candidates. L'approche la plus communément utilisée dans la littérature se base sur des ressources lexicales telles que WordNet. Dans cette approche, les réponses candidates sont écartées si elles ne se trouvent pas dans la hiérarchie de la ressource correspondant au type de réponse attendue par la question [101, 5, 47]. Dans [101], les auteurs proposent un modèle probabiliste de classification de réponses basé sur la régression logistique pour estimer la probabilité qu'une réponse candidate soit correcte.

Précédemment, nous avons passé en revue les systèmes questions-réponses en général. Dans ce qui suit, nous nous intéressons à une sous-catégorie de ces approches : les systèmes question-réponse 5W1H. En effet, comme nous l'avons expliqué dans l'introduction de la section 2.3, la raison pour laquelle nous nous intéressons à ces

approches est que nous les utiliserons dans notre proposition.

2.3.4 Les approches 5W1H

L'approche 5W1H (What, Who, When, Where, Why et How) est souvent utilisée dans la littérature pour extraire des informations structurées sur les événements à partir d'articles de presse en ligne [73, 190]. Par exemple, le titre de l'article de presse suivant répond à cinq des questions 5W1H : Donald Trump a signé un décret ce mardi à la Maison Blanche pour prioriser la livraison de vaccins aux États-Unis. Les phrases surlignées répondent respectivement aux questions Who, What, When, Where et Why.

Les approches proposées dans la littérature pour extraire les réponses aux questions 5W1H à partir d'articles de presse sont généralement basées sur trois étapes [73, 188] :

- **Traitement du texte d'entrée** : les données d'entrée des SQR 5W1H sont souvent constituées d'articles complets, y compris le titre, le paragraphe introductif et le texte principal [157] ou bien un ensemble de phrases (par exemple, les phrases mises en caractères gras) [197]. Ce texte est prétraité en utilisant plusieurs techniques NLP similaires à celles utilisées dans les SQR classiques (section 2.3.3, module de traitement de questions). Par exemple dans [157, 73], les auteurs utilisent des techniques de segmentation de phrase, d'étiquetage grammatical et de reconnaissance d'entités nommées pour analyser les articles de presse. Dans [178], les auteurs utilisent en plus des techniques listées précédemment, des techniques d'analyse grammaticale des dépendances (Dependency Parsing) et de résolution des coréférences (Co-reference Resolution).
- **Extraction de phrases** : cette étape a pour but d'extraire les phrases candidates pour répondre à chacune des questions 5W1H à partir du texte prétraité. Pour ce faire, plusieurs méthodes et stratégies sont proposées dans la littérature [157, 198, 98, 178, 73]. Par exemple, les auteurs dans [157, 198, 98] proposent des méthodes basées sur des règles linguistiques établies manuellement. Dans le système proposé dans [98], les syntagmes nominaux sont identifiés comme candidats *Who* (c.-à-d., les expressions candidates à la réponse à la question *Who*), tandis que les syntagmes verbaux adjacents sont identifiés comme candidats *What*. Dans [178], les auteurs proposent de chercher uniquement les phrases qui ont la structure syntaxique *qui a fait quoi à qui* (par exemple, les manifestants se sont rassemblés pour soumettre un appel au gouvernement) pour en extraire ensuite les candidats *Who* et *What*. D'autres travaux utilisent directement les résultats de la reconnaissance des entités nommées. Par exemple, dans [73], pour déterminer les candidats *Who*, les auteurs proposent de sélectionner toutes les entités nommées qui ont été identifiées comme une personne ou une organisation durant l'étape de traitement. Pour déterminer les candidats *Where* et *When*, ils sélectionnent les entités nommées identifiées comme un lieu, une date ou une heure. Parmi les questions 5W1H, la question *Why* (la cause de l'événement) est la plus compliquée

à répondre étant donné que souvent la raison n'est décrite qu'implicitement dans les articles, voire pas du tout [67]. Dans [67, 6], les auteurs conduisent deux études autour de la détection automatique des relations de causalité et proposent des listes de verbes (ex. "produce", "generate"), d'adverbes (ex. "therefore", "hence") et de phrases de conjonction causale (ex. "consequence of", "for the reason that"). Ces deux études ont constitué la référence de plusieurs travaux pour la réponse à la question *Why*. Par exemple, dans [73], pour identifier les candidats *Why* (pour illustrer nous prenons l'exemple de l'événement du tsunami du Japon en 2011), les auteurs proposent de parcourir l'arbre syntaxique retournée par l'étape de l'étiquetage grammatical et sélectionner les phrases ayant la syntaxe "**syntagme nominal** - **syntagme verbal** - **syntagme nominal**" (par exemple, *les tremblements de terre* - *génèrent* - *les raz-de-marée*). Le syntagme verbal de ces phrases est ensuite analysé pour voir s'il contient un verbe de causalité en se basant sur la liste proposée dans [6]. Dans le cas où le syntagme verbal contient un verbe de causalité, le deuxième syntagme nominal de la phrase est identifié comme candidat *Why* (c.-à-d. *les raz-de-marée*). De même, les auteurs dans [178] proposent une méthode similaire basée sur la liste proposée dans [67] qui est plus récente et plus complète. Des approches plus complexes basées sur des algorithmes d'apprentissage automatique et des ensembles d'entraînement annotés manuellement (un ensemble d'articles de presse avec pour chaque article les bonnes réponses aux questions 5W1H) ont été également proposées dans la littérature [136, 145].

- **Évaluation de candidats** : la dernière étape des systèmes question-réponse 5W1H consiste à ordonnancer les candidats par pertinence. Pour ce faire, plusieurs mesures sont utilisées telles que l'apparition au plus tôt dans le texte [178, 73], la fréquence d'apparition [178, 73] et le nombre de mots du candidat [157]. Quelques travaux proposent des mesures originales, par exemple dans [178], les auteurs proposent une mesure qu'ils appellent *le score de similarité des phrases*. Cette mesure capture l'importance de la phrase dans laquelle le candidat apparaît par rapport à l'ensemble du document. Elle est calculée en se basant sur la technique d'apprentissage automatique "plongement de mots" (Word Embedding) ainsi que sur la similarité cosinus. Des approches plus complexes emploient des algorithmes d'apprentissage automatique [197, 74]. Par exemple dans [197], les auteurs utilisent trois sous-systèmes indépendants pour extraire les réponses. Ensuite, l'algorithme d'apprentissage automatique SVM détermine lequel des trois systèmes retourne la bonne réponse.

Pour résumer, dans cette section nous avons passé en revue les travaux concernant les SQR 5W1H. Nous avons vu que ces systèmes sont généralement basés sur trois étapes principales : (i) traitement du texte d'entrée, (ii) extraction de phrases et (iii) évaluation de candidats. Enfin, nous avons tiré la conclusion que l'approche 5W1H permet de contextualiser la recherche d'information tout en structurant la réponse à l'utilisateur. Ce type d'approche peut donc nous permettre d'améliorer la RI dans les environnements hybrides.

Après avoir exploré les travaux sur la RI classique, la RIS et les SQR pour les corpus

documentaires, dans la section suivante, nous verrons les métriques les plus communément utilisées pour évaluer ces systèmes et nous expliquons ensuite pourquoi nous utilisons certaines de ces métriques dans notre contexte.

2.3.5 Évaluation des systèmes de recherche d'information

L'évaluation d'un système de RI est le processus qui consiste à associer systématiquement une métrique quantitative aux résultats produits par ce dernier en réponse à un ensemble de requêtes d'utilisateurs [21]. Plusieurs métriques d'évaluation ont été proposées [177, 146]. Dans ce qui suit, nous présentons les métriques les plus communément utilisées dans la littérature et nous identifions ensuite celles qui nous paraissent intéressantes dans le contexte de l'explication d'événement.

- **Précision, rappel et F-score** : la précision et le rappel sont des mesures classiques utilisées depuis longtemps dans le domaine de la RI. Le F-score permet de combiner le rappel et la précision. Ces métriques sont définies comme suit :

- **La précision** est la proportion des documents pertinents parmi l'ensemble des documents retournés à l'utilisateur.

$$Précision = \frac{\text{Nombre de documents pertinents retourné}}{\text{Nombre de documents retourné}}$$

- **Le rappel** est le nombre de documents pertinents extraits par rapport au nombre de documents pertinents dans le dépôt de données.

$$Rappel = \frac{\text{Nombre de documents pertinents retourné}}{\text{Nombre de documents pertinents dans le dépôt de données}}$$

- **Le F-score** ou F-mesure combine la précision et le rappel en leur donnant la même importance.

$$F\text{-score} = 2 \cdot \frac{\text{précision} \cdot \text{rappel}}{\text{précision} + \text{rappel}}$$

- **Précision@K ou P@K** : dans les systèmes où l'on retourne un grand nombre de documents (par exemple, la recherche Web), calculer la précision et le rappel sur des milliers de documents retournés n'a plus vraiment de sens puisque l'utilisateur ne va pas tous les regarder et donc on s'intéresse aux documents retournés dans les premières positions. La Précision@K ou P@K correspond au nombre de résultats pertinents parmi les K premiers documents extraits, elle est définie comme suit :

$$Précision@K = \frac{\text{Documents pertinents retourné dans les K premières positions}}{K}$$

- **DCG (Discounted cumulative gain)** : la précision et le rappel, bien que largement utilisés, ne permettent que des évaluations binaires de la pertinence. En utilisant ces métriques, on ne peut pas distinguer entre les modèles de la RI qui retournent des documents très pertinents en haut du classement et ceux qui récupèrent des documents légèrement pertinents en même position. La mesure DCG permet de considérer plusieurs niveaux de pertinence (ex. document pertinent, document moyennement pertinent et document non pertinent). Le DCG à une position de classement p est défini par :

$$DCG_p = \sum_{i=1}^p \frac{2^{rel_i} - 1}{\log_2(i + 1)}$$

Où rel_i est le niveau de pertinence du document au rang i

- **MRR (Mean Reciprocal Rank)** : la mesure MRR teste la faculté d'un système de RI à répondre à un ensemble de requêtes. La mesure MRR favorise le fait que le système fournisse la bonne réponse en première position. Le MRR se calcule en additionnant l'inverse du rang de la réponse la mieux classée : une réponse correcte au premier rang vaut 1 point, une au deuxième vaut 1/2 point et ainsi de suite. Le total est ensuite divisé par le nombre de réponses de façon à obtenir le MRR du système. La mesure MRR est calculée comme suit :

$$MRR = \frac{1}{\text{Nombre de questions}} * \sum \frac{1}{r_i}$$

r_i étant le rang de la bonne réponse pour la $i^{\text{ème}}$ question.

Pour résumer, dans cette section nous avons présenté les métriques les plus communément utilisées dans la littérature pour évaluer les systèmes de RI. Les deux métriques **Précision@K** et **DCG** nous semblent les plus adaptées à notre contexte puisqu'elles nous permettent de savoir si le système retourne les bonnes explications dans les premières places du classement. Cependant, quelques difficultés peuvent être rencontrées, nous les détaillons dans la section suivante.

Après avoir exploré les travaux sur la RI classique, la RI sémantique et les SQR pour les corpus documentaires ainsi que les métriques utilisées pour évaluer ces systèmes, dans la section suivante, nous rappelons tout ce que nous avons vu dans ces parties. Ensuite, nous détaillons les lacunes existantes par rapport à notre contexte. Enfin, nous identifions les éléments qui nous semblent intéressants pour résoudre les défis de cette thèse.

2.3.6 Synthèse et discussion

Dans cette deuxième partie du chapitre état de l'art, nous nous sommes intéressés à la recherche d'information classique (section 2.3.1) et sémantique (section 2.3.2) dans les corpus documentaires, nous avons également exploré les systèmes question-réponse (section 2.3.3) et nous avons mis l'accent sur l'approche 5W1H (section 2.3.4).

Enfin, nous avons passé en revue les métriques d'évaluation des systèmes de recherche d'information documentaire les plus communément utilisées dans la littérature (section 2.3.5).

Les approches classiques et sémantiques de la RI dans les corpus documentaires utilisent des modèles et des techniques différents comme le montre le tableau 2.2. Nous nous sommes focalisés sur les trois étapes : (i) traitement de la requête, (ii) recherche et (iii) présentation des résultats, étant donné que celles-ci sont les étapes communes aux différents processus adoptés par les systèmes de la RI classique et sémantique :

RI corpus de documents	RI classique	RI sémantique	SQR
Traitement de la requête	Mots-clés, phrase en langage naturel, expression booléen Techniques NLP, modification de la requête (ex. sac de mots)	Mots-clés, phrase en langage naturel Techniques NLP, annotation sémantique, théorie des graphes, modification de la requête (ex. désambiguïsation)	Question en langage naturel (ex. questions 5W1H) Techniques NLP, annotation sémantique, théorie des graphes, modification de la requête (ex. désambiguïsation)
Recherche	Appariement exacte (modèle booléen), meilleure appariement (modèles vectoriel, probabiliste et combiné)	Meilleure appariement (modèles vectoriel, probabiliste et combiné), correspondance des graphes, requête SPARQL	Meilleure appariement de passage (modèles de la RI classique et de la RIS adaptés)
Présentation du résultat	Documents non-classés (modèle booléen), documents classés (modèles vectoriel, probabiliste et combinée)	Documents classés	Passages classés

TABLE 2.2 – Techniques adoptées par les approches de la RI classique, de la RIS sémantique et des SQR pour les corpus documentaires

- **Traitement de la requête : la RI classique** pour les corpus documentaires emploie des requêtes mots-clés, des expressions booléennes ou des phrases en langage naturel [160]. **La RI sémantique** a réalisé une avancée majeure comparée aux approches classiques. Elle emploie les techniques du Web sémantique pour améliorer la précision et le rappel des systèmes de la RI. Les requêtes sont exprimées généralement en langage naturel, elles sont analysées et traduites sous une forme sémantique (par exemple, concepts ou triplets RDF) [180, 128] en s'appuyant sur plusieurs techniques (par exemple, techniques NLP, annotation sémantique, etc.).
- **Recherche** : que ce soit en **RI classique** ou en **RIS**, l'étape de recherche consiste à établir l'appariement pour rapprocher la représentation de la requête de celle du corpus documentaire transcrit dans l'index (par exemple, fréquence des mots-clés, graphe sémantique, etc.). Néanmoins, les modèles de **la RIS** diffèrent des approches traditionnelles par la manière dont ils recherchent les ressources. Par exemple, les approches s'appuyant sur les graphes sémantiques emploient de nouvelles techniques de recherche, telles que la traversée de graphes [152] et la correspondance des sous-graphes [180]. Les **SQR** s'appuient sur des systèmes de

la RI classique et de RI sémantique et leurs apportent quelques modifications (par exemple, filtrage de passages au lieu du filtrage de documents).

- **Présentation du résultat** : les approches de **la RI classique** pour les corpus documentaires basées sur le modèle booléen ne fournissent pas de classement pour les ressources récupérées. Les autres approches de **la RI classique** et de **la RI sémantique** produisent à l'étape finale du traitement de la requête une liste classée de documents pertinents qui sont éventuellement accompagnés d'un extrait du passage le plus important du document. **Les SQR** retournent une liste de passages courts pertinents par rapport à la question de l'utilisateur.

Pour résumer, les approches de la RI (**RI classique, RIS et SQR**) conviennent à plusieurs applications, elles diffèrent selon les techniques d'analyse appliquées aux documents et à la requête de l'utilisateur, le temps de traitement et la qualité des résultats retournés. Néanmoins, dans le contexte de l'explication d'événement, il y a encore certaines limites. Tout d'abord, les familles de systèmes que nous avons étudiées sont spécifiquement conçues pour la RI documentaire, par conséquent, **aucun des systèmes que nous avons présentés ne permet à la fois l'interrogation des données de capteurs et des corpus documentaires (défi 1.1., section 1.2.3) ni la combinaison de ces ressources pour répondre à une même requête (défi 1.2., section 1.2.3)**. Les données de capteurs et les données des corpus documentaires ont été toujours perçues comme des sources de données distinctes. Dans ce contexte, l'approche proposée dans les deux travaux [69, 56] qui consiste à utiliser les ontologies pour représenter sémantiquement le contenu de sources de données hétérogènes, nous paraît être un bon point de départ (hypothèse 1., section 1.2.3).

Par ailleurs, **l'approche 5W1H** souvent utilisée dans la littérature pour la structuration des explications d'événements issus des articles de journaux, nous paraît être un excellent choix pour résoudre le défi de la représentation des explications (défi 3, 1.2.3). En effet, que ce soit un événement dans un article de journal ou un événement déclenché dans un environnement connecté, les types de questions que l'on se pose sont toujours les mêmes : pourquoi l'événement a-t-il eu lieu (Why)? Quand est-ce que l'événement a eu lieu (when)? Où est-ce qu'il a eu lieu (Where)? Quelles sont les entités de l'environnement qui ont contribué au déclenchement de cet événement (Who)? et comment (How)? Les six questions 5W1H nous permettent de découvrir tous les aspects liés à l'événement. De plus, nous pensons que la structuration de la réponse sous forme de question-réponse 5W1H assurera l'aspect simple des explications (hypothèse 3., section 1.2.3).

Enfin, concernant les métriques d'évaluation présentées dans la section 2.3.5, la **Precision@K** et le **DCG** sont les métriques les plus adaptées à notre contexte étant donné qu'elles nous permettent de savoir si le système retourne les bonnes explications dans les premières places du classement. Toutefois, ces deux métriques nécessitent de savoir à l'avance les bonnes réponses, ce qui est compliqué dans notre contexte, puisqu'il n'y a pas de système similaire existant dans la littérature, donc très peu ou pas de jeux de données.

Après avoir exploré les systèmes de recherche d'information dans les environnements connectés (section 2.2) et dans les corpus documentaires (section 2.3), dans la section suivante nous passons en revue les méthodes d'interconnexion de données issues d'environnements hybrides. Nous explorons ces travaux puisque nous avons besoin de combiner les informations issues du réseau de capteurs et des corpus documentaires pour pouvoir construire les explications (section 1.2.2).

2.4 Interconnexion de données issues d'environnements hybrides

Les environnements hybrides (ou hétérogènes dans quelques travaux [56, 69]), sont généralement définis comme des environnements recueillant des données de différents formats (par exemple, PDF, TXT, MP3, PNG, etc.) et structures (par exemple, documents structurés, semi-structurés ou non structurés) couvrant divers sujets. Dans notre contexte, nous proposons la définition formelle suivante :

Définition 3. *Un environnement hybride est un environnement connecté regroupant un ensemble de capteurs et de documents, tel que :*

$$EH = \bigcup_{i=0}^n c_i \oplus \bigcup_{j=0}^m d_j \quad \forall i, j \in \mathbb{N}$$

Où :

- **EH** est l'acronyme d'environnement hybride, **c_i** est une instance de capteurs et **d_j** est un document appartenant à EH.

L'interconnexion de sources de données issues des environnements hybrides est une tâche qui a gagné en importance au cours des dernières années. Ceci est dû principalement au progrès du support matériel et au développement de nouvelles technologies, telles que le Big Data ou l'internet des objets (IoT). L'interconnexion de données consiste à **traduire** puis **fusionner** un ensemble de données en un format unique pour pouvoir ensuite les interroger conjointement ou bien les utiliser par des applications pour proposer des services [62]. Récemment, et avec l'avènement des technologies du Web sémantique, la majorité des travaux sur l'interconnexion de données hétérogènes [69, 56, 62, 44, 149, 201, 180], utilisent les ontologies et les graphes RDF comme format final auquel toutes les données sont traduites. Par la suite, pour mettre en correspondance ces représentations et ainsi interconnecter toutes les sources de données, les techniques principales utilisées dans la littérature sont *l'alignement d'ontologies* (ontology matching) et *la liaison de données* (data interlinking) [49]. Dans cette section, nous définissons ce qu'est l'alignement d'ontologies et nous passons en revue les approches principales dans la section 2.4.1. De même, la section 2.4.2 définit le processus de liaison de données et présente les différentes approches. Enfin, nous détaillons dans la section 2.4.3, les stratégies principales de l'interconnexion de données.

2.4.1 Alignement d'ontologies

L'alignement d'ontologies consiste à trouver des correspondances sémantiques entre deux (ou plusieurs) ontologies pour ainsi faire face au problème d'hétérogénéité sémantique. Dans tous les travaux que nous avons consultés et qui définissent formellement ce qu'est un alignement [42, 2, 86], seules les définitions proposées par Euzenat [49] sont à chaque fois reprises. Ce livre est l'une des références bibliographiques les plus populaires et communément citées sur le sujet de l'alignement d'ontologies. Dans ce qui suit, nous détaillons ces définitions et nous présentons quelques exemples.

Définition 4. *Le processus d'alignement est une fonction (f) qui génère un alignement (A_1) à partir de deux ontologies O_1 et O_2 , un alignement initial (A_0), un ensemble de paramètres (P) et un ensemble de ressources (R) :*

$$A_1 = f(O_1, O_2, A_0, P, R)$$

Les trois paramètres (A_0), (P) et (R) sont des paramètres optionnels et ne sont pas toujours utilisés. La Figure 2.5 est la présentation schématique du processus d'alignement. ■

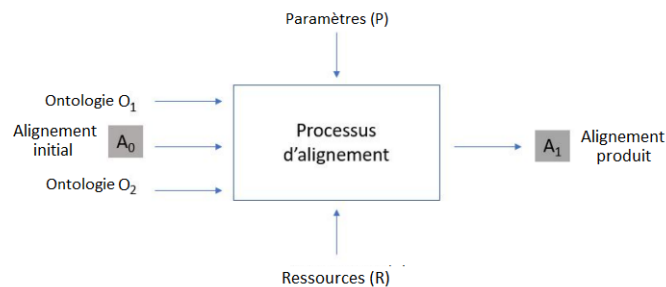


FIGURE 2.5 – Le processus d'alignement²⁰

Définition 5. *Un alignement (figure 2.5) est un ensemble de correspondances $A_{O_1 \rightarrow O_2} = \{c_1, c_2, \dots, c_n\}$ allant d'une ontologie O_1 vers une ontologie O_2* ■

Définition 6. *Une correspondance (ou un mapping) entre deux entités e_1 et e_2 (ex, concept, propriété) appartenant respectivement à deux ontologies O_1 et O_2 , est un quadruple $\langle e_1, e_2, r, s \rangle$ avec :*

- r : la relation entre e_1 et e_2 par exemple l'équivalence (\equiv), généralisation (\supseteq), ou spécification, (\sqsubseteq).
- s : le niveau de confiance de la relation retournée par une mesure de similarité.

■

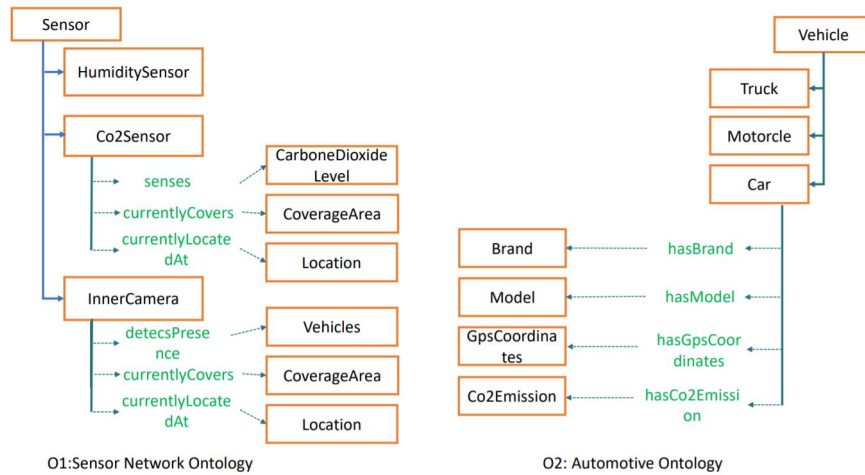


FIGURE 2.6 – Fragments des ontologies réseaux de capteurs et automobile

Un extrait d'une ontologie réseau de capteurs et d'une ontologie automobile sont présentés respectivement à gauche et à droite de la figure 2.6. Les rectangles représentent les concepts. Les propriétés sont représentées en vert. Les correspondances (définition 6) suivantes peuvent être établies :

$$c_1 = (O_1 : \text{Vehicles}, O_2 : \text{Vehicle}, \equiv, 0.71)$$

$$c_2 = (O_1 : \text{Location}, O_2 : \text{GpsCoordinates}, \equiv, 0.73)$$

$$c_3 = (O_1 : \text{currentlyLocated}, O_2 : \text{hasGpsCoordinates}, \equiv, 0.78)$$

Un grand nombre de techniques d'alignement d'ontologies ont été proposées dans la littérature [96, 95, 45, 7, 64, 91, 192, 63, 89, 27, 97]. Plusieurs travaux ont essayé de classer ces techniques sous forme de catégories [10, 168, 77, 49]. Nous nous sommes inspirés de ces travaux pour établir une synthèse qui est représentée dans la figure 2.7. Cette classification est basée sur la manière dont les techniques d'alignement interprètent les informations d'entrée. Par exemple, une approche peut considérer les *labels* des entités des ontologies (concepts et propriétés) comme une *chaîne de caractères* (une séquence de lettres), une autre approche peut les considérer comme des mots dans un langage naturel. Dans ce qui suit, nous passons en revue chacune de ces catégories.

- **Techniques terminologiques** : les techniques terminologiques se basent sur l'analyse des chaînes de caractères, telles que les labels, les commentaires et les descriptions des entités des ontologies pour établir l'alignement [49]. Deux sous-familles principales des techniques terminologiques peuvent être distinguées, à savoir les techniques syntaxiques et les techniques linguistiques. **Les techniques syntaxiques** analysent les chaînes de caractères liées aux entités de plusieurs manières différentes, soit en les considérant comme des séquences de lettres, ensemble de lettres ou ensemble de mots, etc. Selon la manière dont on voit la

20. Figure issue de [49]

21. Synthèse établie à partir des travaux [10, 168, 77, 49]

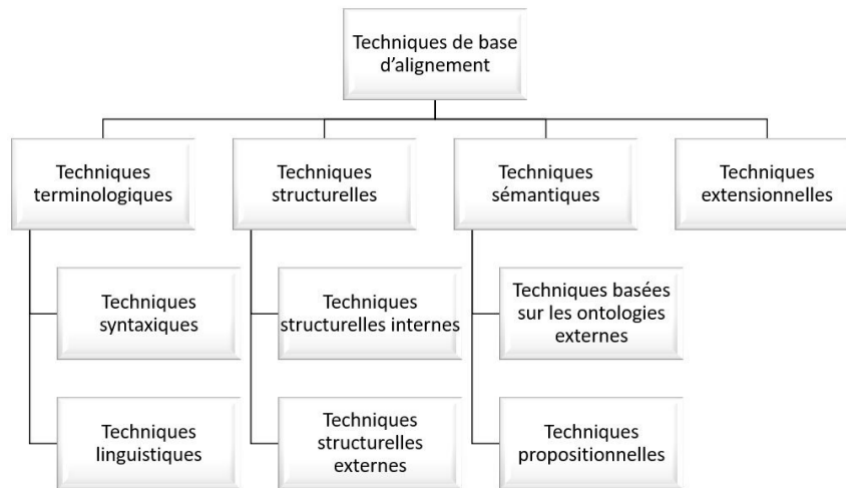


FIGURE 2.7 – Synthèse des techniques d'alignement d'ontologies²¹

chaîne de caractères plusieurs mesures de similarité peuvent être utilisées [29, 29]. Par exemple, les auteurs dans [96, 95] proposent des systèmes d'alignement d'ontologies qui combinent plusieurs mesures de similarité, telles que les distances de levenshtein, Smith-Waterman et Jaro. **Les techniques linguistiques** ne considèrent pas les termes comme des ensembles de lettres mais plutôt comme des mots appartenant à un langage naturel. Pour comparer ces mots plusieurs techniques NLP (ex. lemmatisation, racinisation) et des ressources linguistiques telles que les dictionnaires, les thésaurus, les taxonomies ou d'autres ontologies, sont utilisées pour analyser et calculer la similarité entre les termes. Généralement, les ressources linguistiques sont utilisées pour découvrir les relations sémantiques existantes entre les termes telles que la synonymie, l'hyponymie (spécification) et l'hyponymie (généralisation) ou bien pour mesurer la distance (le chemin de liaison) entre deux mots [49]. Par exemple, dans [45, 7] les auteurs proposent deux systèmes d'alignement d'ontologies qui utilisent l'outil Wordnet pour calculer la similarité entre les entités des ontologies. Les auteurs dans [154, 118], utilisent plutôt des ontologies issues du web.

- **Techniques structurelles** : ces méthodes calculent la similarité entre deux entités en exploitant les informations sur la structure de l'ontologie. Elles se déclinent également en deux sous-catégories : les techniques structurelles internes et les techniques structurelles externes. Pour calculer la similarité entre deux entités d'une ontologie, **les techniques structurelles internes** s'appuient, en plus des labels et des annotations de l'entité, sur des critères tels que par exemple les propriétés de type "DataType" (les propriétés qui relient une classe et un "DataType"), leur domaine et leur portée, leurs caractéristiques (symétrie, transitivité, cardinalité, etc.) et éventuellement les restrictions sur ces propriétés [49]. Par exemple, Glückstad dans l'article [64], propose un système d'alignement d'ontologies basé

sur une combinaison des techniques terminologiques et structurelles internes. **Les techniques structurelles externes** considèrent les ontologies comme étant des graphes étiquetés (les étiquettes étant les labels des entités), la structure externe des entités fait référence à l'ensemble des relations qu'une entité entretient avec d'autres entités [49]. L'intuition derrière ces méthodes est que si deux entités sont similaires alors leur voisinage (ex. entité parents, sous-entités) doit l'être également. Un exemple de ces techniques peut être trouvé dans les deux articles suivants [91, 192].

- **Techniques sémantiques** : ces techniques s'appuient sur des alignements initiaux établis par l'une des techniques présentées dans les sections précédentes. Ces alignements constituent un premier repère sur lequel les méthodes déductives vont s'appliquer afin de vérifier la conformité de ces alignements ou pour en inférer de nouveaux. Les techniques d'alignements sémantiques se distinguent en deux catégories : les techniques basées sur les ontologies externes et les techniques propositionnelles. **Les techniques basées sur les ontologies externes** utilisent des ressources telles que les ontologies de haut niveau ou les ontologies domaines pour déduire les alignements. Une proposition dans ce sens se trouve dans l'article [63]. **Les techniques propositionnelles** s'appuient sur des modèles logiques, tels que par exemple la satisfiabilité propositionnelle (SAT) ou les logiques de description pour déduire les alignements. Deux exemples de systèmes d'alignement basés sur ces techniques peuvent être trouvés dans [89].
- **Techniques extensionnelles** : ces techniques sont basées sur l'intuition que si deux classes partagent un grand ensemble d'instance communes alors il y a de grandes chances qu'elles soient similaires [49]. L'extension désigne l'ensemble des instances d'une classe. Dans le cas où les classes partagent des instances en commun facilement repérables (même URI), des distances telles que la distance de Hamming et la distance de Jaccard sont employées pour calculer la similarité entre les entités en se basant sur l'ensemble des instances. Dans le cas contraire (pas instances en commun facilement repérables), un processus d'identification des instances identiques dans les deux ontologies est conduit. Celui-ci est généralement basé sur des méthodes terminologiques ou structurelles internes citées précédemment. Deux exemples de cette catégorie d'approches peuvent être trouvés dans les articles [27, 97].

Enfin, bien que le problème d'alignement d'ontologies a été largement exploré dans la littérature, l'initiative d'évaluation de l'alignement d'ontologies (OAEI) (Ontology Alignment Evaluation Initiative)²² a prouvé qu'il n'existe pas de méthode ou de système optimal pour tous les problèmes d'alignement existants.

22. <http://oaei.ontologymatching.org/>

2.4.2 Liaison de données

La liaison de données (data interlinking) consiste à créer des liens (principalement des liens *owl:sameAs*) entre les instances de deux ontologies [49]. Peu de travaux ont défini formellement le processus de liaison de données, nous pouvons en citer trois [49, 12, 182]. Ces travaux proposent la même définition formelle suivante :

Définition 7. *La liaison de données vise à trouver un ensemble de liens, principalement des connexions de type *owl:sameAs*, entre deux graphes RDF. L'algorithme d'interconnexion prend en entrée deux graphes RDF d et d' avec un alignement A et génère un ensemble de liens L . Ceci peut être exprimé ainsi :*

$$\text{Interlink}(d, d', A) = L$$

■

Jérôme Euzenat et Pavel Shvaiko [49] proposent également la figure 2.8 qui montre la relation entre le processus d'alignement et le processus de liaison de données. Le processus d'alignement (Matcher, figure 2.8) visent à relier les concepts de deux ontologies (o et o' , figure 2.8), tandis que, le processus de liaison de données (*Linker*, figure 2.8) se base sur l'alignement produit (A , figure 2.8) pour relier les instances des ontologies (d et d' , figure 2.8) avec des connexions de type *owl:sameAs* (L , figure 2.8).

Plusieurs systèmes de liaison de données ont été proposés dans la littérature, les auteurs dans [194] réalisent une revue des divers systèmes de liaison de données proposés dans la littérature. Deux approches principales peuvent être distinguées : les approches basées sur des mesures de similarité et les approches basées sur l'extraction de clés.

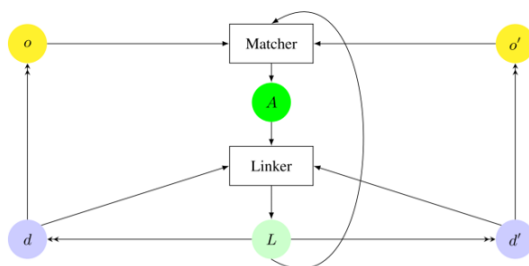


FIGURE 2.8 – La liaison de données et l'alignement d'ontologies²³

- **Les approches basées sur la similarité** : ces approches comparent les graphes au moyen d'une mesure de similarité. Si deux instances sont suffisamment similaires, elles sont considérées comme identiques et le lien *owl:sameAs* est généré. La similarité est basée soit sur les valeurs des instances, soit sur la structure du graphe [194]. Pour comparer les valeurs des instances et la structure du graphe des techniques terminologiques et structurelles internes similaires à celles présentées

23. Figure extraite du livre de Jérôme Euzenat et Pavel Shvaiko [49]

dans la section 2.4.1 sont utilisées. Nous pouvons citer dans cette catégorie, Silk [183] et RDF-AI [156] deux des premiers systèmes d'interconnexion de données proposés dans la littérature ou des systèmes plus récents tels que [196, 30].

- **Les approches basées sur l'extraction de clés** : ces approches spécifient une ou plusieurs propriétés qui permettent d'identifier une instance d'une façon unique (par exemple, ISBN pour les instances de livres similaires, agrégation du nom, prénom et adresse pour identifier de manière unique les individus) [13]. Deux exemples de cette catégorie d'approches peuvent être trouvés dans [13, 12].

En plus de ces deux grandes familles d'approches, récemment plusieurs travaux combinent ces approches classiques avec d'autres techniques telles que l'apprentissage automatique [50, 93] ou la programmation génétique [81].

Après avoir exploré les deux principales techniques d'interconnexion de données, à savoir *l'alignement d'ontologies* et *la liaison de données*, nous verrons dans la section suivante les principales stratégies d'interconnexion de données. Nous essaierons de répondre à la question suivante : à quel moment du traitement le système interconnecte-t-il les données, cela se fait-il entièrement hors ligne, ou bien est-ce que le système interconnecte les données selon les besoins ?

2.4.3 Stratégies d'interconnexion

Nous distinguons deux stratégies principales d'interconnexion de données (que ce soit de l'alignement d'ontologies ou de la liaison de données) dans la littérature, à savoir, (i) l'interconnexion hors ligne et (ii) l'interconnexion en ligne (ou l'interconnexion dynamique).

- **L'interconnexion hors ligne** : c'est la stratégie la plus adoptée dans la littérature [69, 56, 62, 44, 149, 201, 180]. Elle consiste à interconnecter toutes les données à travers la construction d'un grand réseau sémantique. Cette étape se fait en amont, avant l'exploitation de ces ressources par des applications. Des exemples du déploiement de cette approche dans le domaine de la RIS pour interconnecter des ressources hétérogènes (données textuelles et multimédias) sont présentés dans les trois papiers [69, 56, 180]. L'inconvénient de cette approche réside dans la taille du réseau sémantique. Plus la taille des données augmente plus la gestion du réseau et son exploitation devient très coûteuse en termes de temps de traitement et de ressources.
- **L'interconnexion en ligne** : la stratégie d'interconnexion en ligne consiste à conduire le processus d'interconnexion dynamiquement lorsque l'utilisateur ou l'application exprime un besoin d'information. Cette approche est souvent utilisée quand les sources de données sont inconnues au moment de la conception et proviennent de structures autonomes pouvant évoluer (par exemple, les sites Web dont le contenu peut changer constamment)[206]. Un exemple de l'application de l'interconnexion dynamique pour des ressources médicales est présenté dans le papier [206]. Un autre exemple pour l'interconnexion de bases de connaissances

Interconnexion de données	Alignement d'ontologies	Liaison de données
Données d'entrée	Ontologies	Instances d'ontologies
Techniques	Techniques terminologiques, sémantiques, structurelles et extensionnelles	Techniques terminologiques et structurelles
Connections générées	Relations d'équivalence, généralisation et spécification	<i>owl:sameAs</i>
Stratégies	Interconnexion hors ligne, interconnexion en ligne	Interconnexion hors ligne, interconnexion en ligne

TABLE 2.3 – Techniques adoptées par les approches d'interconnexion de données

hétérogènes est présenté dans [129]. L'inconvénient de cette approche réside dans le temps du traitement de la demande d'information, vu que l'interconnexion se fait entièrement après la réception de la demande, si la taille du jeu de données est considérable le temps de traitement le sera aussi.

2.4.4 Synthèse et discussion

Dans cette section, nous avons passé en revue les techniques principales d'interconnexion de données, à savoir, *l'alignement d'ontologies* (section 2.4.1) et *la liaison de données* (section 2.4.2). Ensuite nous avons présenté les stratégies principales d'interconnexion de données (section 2.4.3). Nous récapitulons tout ceci dans le tableau 2.3.

L'alignement d'ontologie établit des relations d'équivalence, de généralisation et de spécification entre les entités des ontologies en utilisant plusieurs techniques terminologiques, sémantiques, structurelles et extensionnelles. La liaison de données quant à elle, établit des relations *owl:sameAs* entre les instances d'ontologies en utilisant des techniques terminologiques et structurelles. Deux stratégies principales d'interconnexion sont présentées dans la littérature : l'interconnexion hors ligne et l'interconnexion en ligne. L'interconnexion hors ligne consiste à interconnecter l'ensemble des données en amont du traitement, tandis que, l'interconnexion en ligne interconnecte les données au moment du traitement.

Dans notre contexte, nous utilisons les ontologies pour représenter les données de capteurs et les documents (hypothèse 1., section 1.2.3), les techniques d'alignement d'ontologies représentent donc un bon point de départ pour interconnecter ces ontologies. Maintenant, la question qui se pose est la suivante : quelle approche adopter pour les deux techniques? En effet, vu que les ontologies dont nous disposons modélisent les données de capteurs et les documents, les concepts et les relations couvrent des domaines d'application hétérogènes. Par conséquent, des approches telles que *les techniques structurelles* qui se basent sur le voisinage des entités qui est très différent

dans notre cas, ou bien, *les techniques extensionnelles* qui nécessitent un grand nombre d'instances ne donneront pas de bons résultats. Nous choisissons donc d'opter pour *les techniques sémantiques basées sur des ontologies externes*, vu que celles-ci s'appuient sur des ressources externes (ex. Wordnet, Wikipédia), qui permettront de rapprocher plus facilement les différentes ontologies. Concernant, la liaison de données, les relations *owl:sameAs* permettront d'interconnecter, dans un premier temps, les ontologies au niveau des instances. Néanmoins, nous devons adapter ces techniques à notre contexte ou bien, carrément ajouter une couche supplémentaire de relations dédiées spécifiquement à l'explication des événements.

Enfin, concernant les stratégies d'interconnexion, l'interconnexion hors ligne du réseau sémantique au niveau des concepts et au niveau des instances (section 2.4.3), pourrait rendre le réseau sémantique beaucoup trop dense. Les opérations de recherche sur ce réseau seront coûteuses en matière de ressources et de temps de traitement. D'un autre côté, attendre que l'événement se déclenche pour lancer le processus d'interconnexion ne semble pas non plus être un bon choix, vu que cela pourrait nécessiter un temps de traitement important. L'utilisateur pourrait avoir besoin d'une explication dans l'urgence (alerte incendie, alerte haut niveau de CO₂, etc.). Une combinaison des deux stratégies à travers **l'alignement hors ligne et la liaison de données en ligne** nous paraît donc être un bon compromis.

2.5 Bilan

Nous avons parcouru dans ce chapitre trois axes de recherche : (i) **la RI dans les environnements connectés** (section 2.2), (ii) **la RI dans les corpus documentaires** (section 2.3) et (iii) **l'interconnexion de données dans les environnements hybrides** (2.4). Dans les deux premiers axes, nous avons détaillé les différentes approches classiques et sémantiques de la RI, respectivement dans les environnements connectés et dans les corpus documentaires. Dans le troisième axe, nous avons présenté les principales techniques utilisées dans la littérature pour l'interconnexion de données dans les environnements hybrides, à savoir l'alignement d'ontologies et la liaison de données. Ensuite, nous avons passé en revue les différentes stratégies d'interconnexion. Dans ce qui suit, nous rappelons les deux défis auxquels nous sommes confrontés dans cette thèse et présentons les pistes de solutions identifiées au regard de ce que nous avons vu dans ce chapitre.

- **Défi 1 : Recherche d'information dans des ressources de données hétérogènes** : comme nous l'avons expliqué dans le chapitre introduction (section 1.2.3), ce défi est composé de deux sous-défis :
 - **Défi 1.2. Interrogation automatisée de différentes sources de données** : pour construire des explications aux événements déclenchés dans un environnement connecté, la combinaison des données de capteurs et de documents est cruciale (section 1.2.2). Toutefois, bien que les modèles actuels de la RI (RI traditionnelle, RIS et SQR) permettent de rechercher des informa-

tions pertinentes contenues dans différents types de ressources, très peu de travaux se sont intéressés à l'interrogation à la fois des données de capteurs et des corpus documentaires. Les données de capteurs et les corpus documentaires ont été toujours perçus comme des sources de données distinctes. Pour résoudre ce problème, nous faisons l'hypothèse que les ontologies nous permettront de décrire les deux sources de données en utilisant un format unique. Ainsi, nous pourrions les interroger conjointement. Dans ce contexte, et comme nous l'avons expliqué dans la section 2.2.3, nous avons choisi d'utiliser l'**ontologie HSSN** [116] parce qu'elle couvre tous les aspects des EC et des capteurs dont nous aurons besoin. Pour modéliser les documents, nous ne pouvons pas, à ce stade, choisir les ontologies que nous allons utiliser parce que nous ne connaissons pas à l'avance leur contenu.

- **Défi 1.2. Combinaison des résultats issus de requêtes multiples** : pour répondre à une requête d'un utilisateur, la majorité des systèmes de la RI ont pour but de trouver la ressource pertinente (une entrée dans une base de données, un document, un passage dans un document), ensuite la retourner à l'utilisateur. Cependant, dans notre contexte, les éléments constituant la réponse à la requête de l'utilisateur (pourquoi l'événement s'est-il déclenché?) sont dispersés dans plusieurs ressources (plusieurs données de capteurs et plusieurs documents). Il est nécessaire de combiner ces ressources pour construire une réponse. Les systèmes de la RI classique et de la RI sémantique retournent plusieurs éléments d'information mais ne les combinent pas. C'est à l'utilisateur de faire ce travail par lui-même. Pour résoudre ce problème, nous avons choisi d'utiliser (1) les techniques **d'alignement d'ontologies** basées sur les ontologies externes pour interconnecter les ontologies au niveau des concepts (les raisons de ce choix sont expliquées dans la section 2.4.4), et (2) les techniques de **liaison de données** pour interconnecter les ontologies au niveau des instances. Bien évidemment, il faudra adapter ces techniques à notre contexte pour que les connexions construites soient au service de la recherche d'explication. Enfin, concernant la stratégie d'interconnexion, nous avons choisi d'adopter un compromis entre l'interconnexion hors-ligne et en ligne des données. Concrètement, nous établissons l'**alignement d'ontologies hors-ligne** et nous procédons à **la liaison de données en ligne**.
- **Défi 2. Présentation d'une réponse bien structurée** : les modèles actuels de la RI présentent le résultat d'une requête de différentes façons (documents classés ou non classés, tableau avec un ensemble de données, etc.). Cependant dans notre contexte, nous devons présenter à l'utilisateur des pistes d'explications d'événements qui incluent plusieurs aspects spatial, temporel, des entités de l'environnement qui ont chacun leur rôle, des documents qui nous ont aidés à relier ces entités à l'événement en question, etc. De simples passages de documents ne seront pas suffisants pour présenter les explications d'une manière claire et simple. Pour résoudre ce problème, **la technique 5W1H** présentée dans la section

2.3.4, nous paraît être une approche intéressante. Les six questions 5W1H (*What, Who, When, Where, How, et Why*) nous permettront de couvrir tous les aspects liés à l'événement (ex. le temps, le lieu, les entités, etc.). Ainsi, nous faisons l'hypothèse que la structuration de la réponse sous forme de question-réponse 5W1H assurera l'aspect simple et clair des explications.

Chapitre 3

Le système ISEE

3.1 Introduction

Dans le chapitre 2, nous avons présenté l'état de l'art autour de trois axes : la RI dans les environnements connectés, la RI dans les corpus documentaires et l'interconnexion des données hétérogènes. Nous avons analysé les différentes approches proposées et nous avons identifié les éléments qui nous semblent intéressants dans le contexte de l'explication d'événements. Dans ce chapitre, nous présentons notre proposition, ISEE (Information System for Event Explanation), un système pour l'explication d'événements dans les environnements hybrides (les environnements connectés impliquant des SI hétérogènes pour la gestion des données de capteurs et des corpus documentaires).

Le système ISEE utilise tous les éléments intéressants précédemment identifiés dans la littérature (chapitre 2) et les adapte à notre contexte, à savoir, **l'ontologie HSSN** [115], **l'approche 5W1H** [73, 190], **les techniques d'alignement d'ontologies** [49] et **les techniques de liaison de données** [49]. D'une part, ISEE est basé sur deux modèles, un modèle pour la définition d'événements (**contribution 1.1**, section 1.3.1) et un modèle pour la définition d'explications d'événements (**contribution 1.2**, section 1.3.1) dans les environnements hybrides. Le modèle pour la définition d'événements étend **l'ontologie HSSN** [116], tandis que le modèle pour la définition d'explication d'événements s'inspire de **l'approche 5W1H** et l'adapte à notre contexte. D'autre part, le système ISEE fournit un processus pour l'interconnexion de réseaux de capteurs et de corpus documentaires (**contribution 2.1**, section 1.3.2). Ce dernier s'appuie sur des **techniques d'alignement d'ontologies** [49] et sur le principe de **liaison de données** pour proposer un nouveau type d'interconnexions sensibles au contexte. Ces interconnexions ont pour but de rapprocher les données de capteurs et le corpus documentaire au service de la construction d'explications. Enfin, ISEE fournit des algorithmes pour récupérer et noter les informations pertinentes afin d'expliquer le déclenchement d'un événement (**contribution 2.2**, section 1.3.2). **L'approche 5W1H** est le principe directeur de notre proposition pour établir des connexions sémantiques entre les données de capteurs et les documents, puis construire les explications potentielles de l'occurrence d'un

événement.

Ce chapitre est organisé comme suit. Dans la section 5.2, nous présentons une vue d'ensemble du système ISEE, nous détaillons brièvement les différents modules et leurs fonctionnalités. La section 3.3 présente les définitions formelles des éléments essentiels de l'environnement. La section 3.4 se focalise sur la modélisation d'événements. Nous présentons tout d'abord, le modèle pour la définition d'événements dans les environnements hybrides (contribution 1.1, section 1.3.1). Ensuite, nous présentons le modèle pour la définition d'explication d'événements (contribution 1.2, section 1.3.1). La section 3.5 détaille le processus d'explication d'événements (contributions 2.1 et 2.2, section 1.3.2). Enfin, nous concluons ce chapitre dans la section 3.6.

3.2 Vue d'ensemble

Comme nous l'avons expliqué précédemment, notre environnement se compose de deux systèmes d'information : un SI de réseau de capteurs et un SI de corpus de documents hétérogènes. Nous supposons que nous disposons de plusieurs ontologies de domaine pour modéliser la sémantique du réseau de capteurs (ontologie HSSN [116], section 2.2.2) et du corpus de documents (par exemple, une ontologie ressources humaines et une ontologie bâtiment). Notre stratégie repose sur une interconnexion classique hors-ligne, suivi de processus itératifs d'interconnexion et de filtrage sensibles au contexte (les données issues de la définition et du déclenchement de l'évènement).

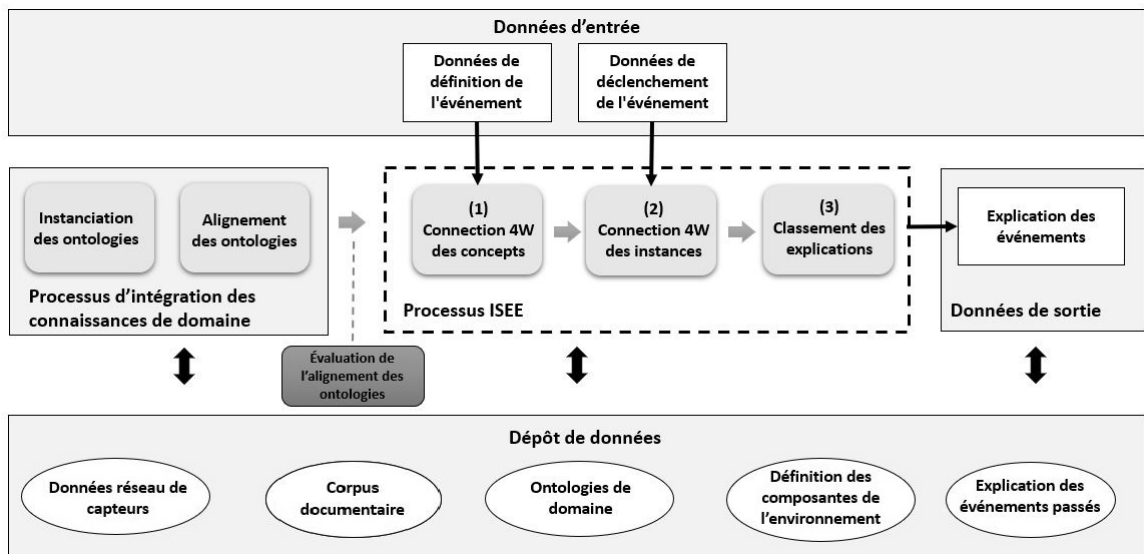


FIGURE 3.1 – Vue d'ensemble du système ISEE

Le système ISEE (figure 3.1) consiste donc en deux processus principaux :

- **Processus d'intégration des connaissances de domaine** : ce processus a pour but (i) d'intégrer toutes les informations provenant de l'environnement (données de

capteurs et corpus documentaire) dans les ontologies de domaine sous forme d’instances, ensuite, (ii) d’interconnecter ces ontologies, dans un premier temps, en utilisant des techniques classiques d’alignement. Le processus d’intégration des connaissances de domaine est donc composé de deux sous-étapes :

- **Instanciation des ontologies** : des techniques d’instanciation automatique des ontologies [138, 38, 141, 110] sont utilisées pour analyser le corpus de documents et ensuite alimenter les ontologies de domaine avec un ensemble d’instances (*instanciation des ontologies* dans la figure 3.1, section 3.5.1.1).
- **Alignement des ontologies** : des techniques d’alignement d’ontologies [42, 2, 86] sont utilisées afin de construire des relations d’alignement classiques (relations d’équivalence, de généralisation et de spécialisation) entre les concepts des ontologies de domaine (*alignement des ontologies* dans la figure 3.1, section 3.5.1.2).

De toute évidence, si de nouveaux documents ou de nouveaux capteurs viennent s’ajouter dans le système d’information, l’étape d’instanciation des ontologies doit être réexécutée pour que les nouvelles instances soient intégrées dans les ontologies de domaine.

Après l’exécution du *processus d’intégration des connaissances de domaine*, le système ISEE peut commencer à construire les explications à travers le *processus ISEE* (*processus ISEE*, figure 3.1). Toutefois, avant d’entamer cette étape, le système offre à l’utilisateur la possibilité d’évaluer l’alignement (***évaluation de l’alignement des ontologies*** dans la figure 3.1, section 3.5.2). Cette étape est facultative, elle permet à l’utilisateur d’obtenir un score qui lui indique si l’alignement établi lui permettra éventuellement d’obtenir de bonnes explications. Ce score est calculé au moyen d’une formule qui sera détaillée dans la section 3.5.2.

- **Processus ISEE** : ce processus s’appuie sur le graphe de connaissances (un ensemble d’ontologies de domaine instanciées et alignées) construit dans l’étape précédente et sur les données de définition et de déclenchement de l’événement pour établir une deuxième couche d’interconnexions sensibles au contexte entre les ontologies de domaine. Le graphe de connaissances est ensuite analysé pour construire les candidats d’explication de l’événement que nous appelons dans le reste de ce manuscrit *les tuples Why*. Le processus ISEE est composé de trois étapes :

(1) **Connexion 4W des concepts** : un premier réseau d’*interconnexions sensibles au contexte* au niveau conceptuel est construit sur la base de la définition de l’événement et des alignements classiques précédemment construits ((1) dans la figure 3.1, section 3.5.3.1).

(2) **Connexion 4W des instances** : sur la base des données de déclenchement d’événements et des résultats de processus de liaison de données, un deuxième niveau d’*interconnexions sensibles au contexte* entre les instances est construit ((2) dans la figure 3.1, section 3.5.3.2).

(3) Classement des explications : le graphe de connaissances est analysé, les candidats d'explication sont construits et classés en fonction d'un score de pertinence ((3) dans la figure 3.1, section 3.5.3.3).

Pour résumer, les avantages de cette approche résident dans le fait qu'elle est :

- **Générique** : du moment où l'on dispose d'ontologies pour modéliser la sémantique du domaine, le système ISEE peut être utilisé dans tout type d'environnement connecté (par exemple, parking, entreprise ou hôpital).
- **Flexible** : le système ISEE est basé sur des outils existants (outils d'instanciation, d'alignement et de liaison de données), l'utilisateur a donc la possibilité de choisir les outils qui lui conviennent pour obtenir de meilleurs résultats. Par exemple, si l'outil d'alignement ne parvient pas à établir les connexions nécessaires à l'explication de l'événement déclenché, l'utilisateur peut choisir un autre outil plus adapté.
- **Optimisée** : le processus d'intégration des connaissances de domaine et le processus ISEE sont exécutés en deux temps : hors ligne avant le déclenchement de l'événement (le processus d'intégration des connaissances de domaine), puis en ligne lorsque l'événement se déclenche et que l'utilisateur souhaite avoir une explication (processus ISEE). Cela permet de réduire le temps de traitement de la demande de l'utilisateur. De plus, les étapes d'interconnexion et de filtrage (*connexion 4W des concepts* et *connexion 4W des instances*, processus ISEE) sont guidées par les données provenant de l'événement. En d'autres termes, nous n'interconnectons pas l'ensemble du réseau sémantique, mais plutôt la sous-partie qui nous intéresse, c'est-à-dire celle en lien avec l'événement. Ces deux stratégies permettent de gagner en terme de ressources et en terme de temps de traitement.

Notez que, dans la figure 4.1, une flèche relie le données de sorti au dépôt de données, cette dernière fait référence au fait que les explications construites par le système sont archivées dans le dépôt de données. Ces explications peuvent éventuellement être utilisées pour améliorer les performances du système, cette partie sera abordée dans des travaux futurs. La manière dont les données sont récupérées (les deux flèches reliant le processus d'intégration des connaissances de domaine et le processus ISEE au dépôt de données) est détaillée dans l'annexe A. Enfin, nous rappelons ici que **les étapes (1) et (2)** du processus ISEE constituent **la contribution 2.1** (section 1.3.2) de cette thèse, tandis que, **l'étape (3)** constitue **la contribution 2.2** (section 1.3.2).

Après avoir présenté une vue d'ensemble du système ISEE, nous explorons dans la suite de ce chapitre chacune de ses composantes. Dans la section suivante, nous listons les différents éléments constituant l'environnement et nous présentons leurs définitions formelles respectives (*définition des composantes de l'environnement*, figure 3.1).

3.3 Définition des composantes d'un environnement hybride

La toute première étape à réaliser avant que le système ISEE puisse commencer à construire des explications est celle de la représentation sémantique des différentes composantes de l'environnement. Lorsqu'un nouveau capteur est mis en place ou lorsqu'une nouvelle entité (par exemple, une voiture ou une personne) entre dans l'environnement, les concepts correspondants dans l'ontologie de réseaux de capteurs et dans les ontologies de domaine sont instanciés automatiquement par le système (détection automatique des nouvelles entités de l'environnement [191, 87, 163]), ou par l'utilisateur via un langage de requête [115]. De même, lorsqu'un nouveau document est ajouté dans le corpus, les concepts des ontologies de domaine correspondantes sont automatiquement instanciés [138, 38, 141, 110]. Pour présenter ces éléments dans notre contexte, nous proposons les définitions formelles suivantes : pour les capteurs (déf. 8), pour les entités (déf. 9) et pour les documents (déf. 10).

Définition 8. Nous définissons un **capteur** comme un tuple de 9 attributs.

$$\langle id, description, mobility, connectivity, A, L, F, Doc, O \rangle$$

- **id** est l'identifiant unique du capteur.
- **description** est la description textuelle du capteur.
- **mobility** désigne la capacité de mobilité du capteur (statique, mobile).
- **connectivity** désigne la manière dont le capteur est accessible (wifi, ethernet, radio, bluetooth, etc).
- **A** est l'ensemble des zones couvertes par le capteur, $A = \{area_1, \dots, area_q\}$ avec $area_{i \in 1, q} = \{\alpha, \beta, distance\}$, $\alpha, \beta \in [0; 2\pi]$ sont les angles qui définissent respectivement les écarts horizontaux et verticaux de la zone de couverture et *distance* est la portée du capteur qui définit l'étendue de la zone de couverture en mètre.
- **L** est l'ensemble des emplacements du capteur, $L = \{area_1, \dots, area_n\}$ avec $area_i = \langle t_i, l_i \rangle$, t_i est l'heure et l_i la localisation du capteur (ex. coordonnées GPS), si le capteur est statique, L ne contiendra qu'une seule localisation $L = \{location\}$.
- **F** désigne l'ensemble des caractéristiques capturées par le capteur $F = \{f_1, \dots, f_s\}$.
- **Doc** désigne l'ensemble des id des documents liés au capteur, $Doc = \{doc_1, \dots, doc_n\}$.
- **O** désigne l'ensemble des IRI (Internationalized Resource Identifier) des concepts liés au capteur. ■

Nous présentons ci-dessous un exemple de la définition d'un capteur de lumière *light_sensor1* :

<"light_sensor1", "IR Dual Beam Sensing Technology ... ", "static", { $2\pi, \pi, 3m$ }, {"Office124"}, {"luminosity"}, {"doc15", "doc124"}, {"https://exp.com/resource.txt#luminosity"}, "https://exp.com/resource.txt#Lumens", "https://exp.com/resource.txt#Lamp"}>

Comme vous pouvez le voir, "light_sensor1" constitue l'identifiant du capteur (attribut *id*). Il est suivi d'une description détaillée qui provient de sa fiche technique (attribut *description*), puis de la valeur "static" qui indique que le capteur a une mobilité statique et ne peut donc pas changer d'emplacement (attribut *mobility*). La valeur " $\{2\pi, \pi, 3m\}$ " indique que le capteur couvre une distance de 3 mètres avec des écarts horizontaux et verticaux de 2π et π respectivement (attribut *A*). Le capteur est situé dans la localisation 'Office124' (attribut *L*). Il capture les niveaux de lumière "luminosity" (attribut *F*). Les deux documents "doc15" et "doc124" sont ceux liés au capteur (attribut *Doc*). Enfin, les trois concepts *Luminosity*, *Lumens* et *Lamp* sont liés au capteur "light_sensor1" (attribut *O*).

En plus des capteurs, l'environnement comporte également des objets que nous appelons des entités. Nous supposons qu'une entité peut être statique ou mobile. Les entités statiques sont celles qui ne peuvent pas changer de position (par exemple, une fenêtre, un climatiseur, un arbre, etc.). Les entités mobiles sont celles qui peuvent disparaître de l'environnement ou changer d'emplacement au fil du temps (par exemple, une voiture, une personne, un téléphone portable, etc.).

Définition 9. Nous définissons une **entité** comme un tuple de 6 attributs.

<id, description, mobility, L, Doc, i>

- **id** est l'identifiant unique de l'entité.
- **description** est la description de l'entité.
- **mobility** désigne la capacité de mobilité de l'entité (statique, mobile).
- **L** est l'ensemble des emplacements de l'entité, il nous permet de garder la trace des entités qui ont la capacité de se déplacer et de quitter l'environnement $L = \{area_1, \dots, area_n\}$ avec $area_i = \langle t_i, l_i \rangle$, t_i est l'heure et l_i la localisation. Si l'entité est statique, *L* ne contiendra qu'une seule localisation $L = \{location\}$.
- **Doc** désigne l'ensemble des id des documents liés à l'entité, $Doc = \{doc_1, \dots, doc_n\}$.
- **i** désigne l'IRI de l'instance qui correspond à l'entité. ■

Nous présentons ci-dessous un exemple de la définition d'une entité mobile de type ordinateur portable Asus Zenbook 13UX325E :

<"laptop_249", "model : Asus Zenbook 13UX325E ... ", "mobile", {<"Office124", "12-10-2019 08 :02 :12">, <"MeetingRoom145", "12-10-2019 12 :03 :22">, ... }, {"doc875"}, "https://exp.com/resource.txt#laptop_249">

"laptop_249" constitue l'identifiant de l'entité (attribut *id*). Il est suivi d'une description détaillée qui provient de sa fiche technique (attribut *description*), puis de la valeur "mobile" qui indique le type de mobilité de l'entité (attribut *mobility*). Ensuite, l'historique des emplacements de l'entité est détaillé (attribut *L*) : l'emplacement "Office124" à "12-10-2019 08 :02 :12" puis l'emplacement "MeetingRoom145" à "12-10-2019

12 :03 :22", etc. La valeur "doc875" indique l'identifiant du document lié à l'entité (attribut *Doc*). Enfin, l'IRI "https://exp.com/resource.txt#laptop_249" est celui de l'instance correspondante à l'entité (attribut *i*).

Définition 10. Nous définissons un **document** comme un tuple de 4 attributs.

$\langle id, type, path, I \rangle$

- **id** est l'identifiant unique du document.
- **type** indique le type de document : texte, image, vidéo ou composite (par exemple, un fichier PPT contenant du texte et des images).
- **path** est l'URL du document.
- **I** dénote l'ensemble des instances mentionnées dans le document avec leurs concepts associés et positions correspondantes dans le document. La structure de cet élément dépend du type de document, par exemple, pour les documents texte $I = \{ \langle instance_URI_1, concept_URI_1, \{ \langle offset_start, offset_end \rangle, \dots, \langle offset_start, offset_end \rangle \} \rangle, \dots \}$, *offset_start* et *offset_end* sont les positions de début et de fin de la séquence de mots faisant référence au concept dans le document. Le décalage est la distance par rapport au début du document qui est mesurée par le nombre de caractères.

■

Nous présentons ci-dessous un exemple de la définition de la fiche technique de l'ordinateur portable Asus Zenbook 13UX325E :

```
<"doc875", "text", "file:///C:/Documents/doc875.pdf", {"https://exp.com/resource.txt#laptop_249", "https://exp.com/resource.txt#Laptop", {<45,48>, <281,284>, ...}>,...}>
```

"doc875" est l'identifiant du document (attribut *id*). La valeur "text" indique que c'est un document texte (attribut *type*), tandis que "https://exp.com/resource.txt#laptop_249" représente l'URL du document (attribut *path*). Enfin, la dernière partie de la définition détaille l'ensemble des instances mentionnées dans le document (attribut *I*), par exemple, "laptop_249" est l'instance du concept "Laptop" qui apparaît plusieurs fois dans le document "doc875", de l'offset 45 à 48 et de l'offset 281 à 284, etc.

Après avoir défini les éléments essentiels de l'environnement, dans la section suivante, nous décrivons nos deux modèles pour la définition d'événements et pour la définition d'explication d'événements. Nous rappelons que ces deux modèles constituent respectivement les contributions 1.1 et 1.2 de cette thèse (section 1.3.1).

3.4 Modélisation des événements dans les environnements hybrides

Comme nous l'avons expliqué précédemment, nous avons choisi l'ontologie HSSN [115] pour modéliser les données de capteurs (sections 2.2.3 et 2.5). Cependant, cette ontologie ne permet pas de modéliser les événements, en d'autres termes, elle ne contient

pas les concepts et les propriétés permettant d'intégrer dans l'ontologie les données de définition et de déclenchement d'événements. Dans notre contexte, il est nécessaire de relier les informations de l'événement avec les informations contenues dans les autres ontologies pour pouvoir ensuite analyser ce graphe de connaissance et construire les explications. L'intégration des données de définition et de déclenchement de l'événement dans l'ontologie de réseaux de capteurs est donc cruciale pour la construction des explications.

Par ailleurs, la modélisation d'événements comporte un autre aspect très important celui de la modélisation des explications. Les explications retournées par le système doivent être structurées de façon simple, intuitive et facile à comprendre pour les utilisateurs des environnements connectés (défi 3, section 1.2.3). Nous avons identifié dans le chapitre 2 l'approche 5W1H comme piste intéressante, nous montrons dans la suite de cette partie comment nous l'avons adaptée à notre contexte.

Dans la section 3.4.1, nous présentons notre modèle multidimensionnel pour la définition d'événements dans les environnements hybrides (contribution 1.1, section 1.3.1). La section 3.4.2 détaille notre modèle inspiré de l'approche 5W1H pour la structuration des explications d'événements.

3.4.1 Un modèle multidimensionnel pour la définition d'événements dans les environnements connectés

Nous définissons un événement comme un espace à 4 dimensions appelé *eSpace*. L'*eSpace* se compose des dimensions *Feature*, *Source*, *Time*, et *Location* ainsi que des données de l'événement pour représenter les observations des capteurs qui ont permises de le détecter. Dans ce qui suit, nous définissons d'abord un événement, puis nous détaillons la définition d'une dimension, enfin, nous définissons l'*eSpace* de l'événement.

Définition 11. Un événement e est défini comme un tuple de 4 attributs.

$$\langle id, l, eSpace, u \rangle$$

- **id** est l'identifiant unique de l'événement.
- **l** est un label.
- **eSpace** est un espace multidimensionnel qui définit les composants de l'événement sous forme de plusieurs dimensions (définition 13) en plus des données liées à la détection de l'événement.
- **u** est le niveau d'urgence attribué à e , où $u \in \{low, medium, high\}$. ■

Par exemple, un événement de gaspillage de lumière peut être défini comme suit : *LightWastageEvent*: $\langle 1, 'LightWastage', eSpace_{LightWastage}, 'low' \rangle$

Le niveau d'urgence indique ici le niveau requis de l'explication de l'événement en fonction de la préférence de l'utilisateur. Par exemple, si $u = low$ (c.-à-d., que l'utilisateur est en mesure d'attendre un certain temps avant de recevoir l'explication), il est

possible d'expliquer de manière complète l'événement, même si le temps de traitement est important. Cependant, si $u = high$ (c.-à-d., il y a un besoin urgent d'explication), l'explication devrait être succincte afin de renvoyer des résultats rapides à l'utilisateur.

Définition 12. Une **dimension d'événement** d est définie sous forme d'un tuple de 5 attributs :

$$\langle id, o, C, R, C_d \rangle$$

- **id** est l'identifiant unique de la dimension.
- **o** est le concept d'origine qui représente le mieux d .
- **C** est l'ensemble des concepts liés à d .
- **R** est l'ensemble des propriétés liées à d .
- **C_d** est l'ensemble des contraintes associées à d . ■

Par exemple, supposons une dimension *Feature* qui traite de la caractéristique de l'événement gaspillage de lumière et une dimension *Time* qui traite de l'aspect temporel de cet événement. Les deux dimensions peuvent être définies comme suit :

Feature_{LightWastage} : $\langle 3, Luminosity, \{Cost, Energy\}, \{hasFeature\}, \{Lumens > 30\} \rangle$

Time_{LightWastage} : $\langle 5, TimeInterval, \{TemporalEntity, TimeInstant\}, \{inDateTime, inTemporalPosition\}, \{timestamp \textit{duration} 1hour, timestamp \textit{includedIn} [07:00:00pm, 06:00:00am]\} \rangle$

Comme vous pouvez le constater, les deux dimensions ont des attributs différents. La dimension *Feature_{LightWastage}* a pour origine (attribut o) le concept *Luminosity* qui décrit la caractéristique de l'événement (niveau de lumière). *Cost* et *Energy* représentent des concepts liés à la dimension *Feature_{LightWastage}* (attribut C). *hasFeature* est une propriété qui est également liée à cette dimension (attribut R). Une seule condition est associée à cette dimension (attribut C_d) : "*Lumens > 30*" pour indiquer que les niveaux de lumière doivent être supérieurs à 30 lux. La dimension *Time_{LightWastage}* a pour origine (attribut o) le concept *TimeInterval* qui décrit l'aspect temporel de l'événement (l'événement est détecté dans un intervalle de temps). *TemporalEntity* et *TimeInstant* représentent des concepts liés à la dimension *Time_{LightWastage}* (attribut C). *inDateTime* et *inTemporalPosition* sont des propriétés également liées à cette dimension (attribut R). Deux conditions sont associées à cette dimension (attribut C_d) : "*timestamp duration 1hour*" et "*timestamp includedIn [07:00:00pm, 07:00:00am]*". "*timestamp duration 1hour*" indique que les conditions de déclenchement doivent être validées durant une heure, tandis que "*timestamp includedIn [07:00:00pm, 06:00:00am]*" indique que celles-ci doivent être validées entre 19h et 06h.

Notez ici que des opérateurs qualitatifs et quantitatifs sont utilisés pour définir des contraintes sur ces dimensions [115]. Par exemple, *includedIn* et *duration* sont des opérateurs dédiés aux données temporelles (c.-à-d., des opérateurs qualitatifs) alors que $>$ est un opérateur pour les données numériques (c.-à-d., un opérateur quantitatif).

Maintenant que nous avons défini la notion de dimension, nous pouvons définir ce qu'est un espace d'événement (*eSpace*) puisque ce dernier inclut des dimensions.

L'Espace d'un événement est constitué de plusieurs dimensions. Cependant, quatre dimensions sont obligatoires par défaut, à savoir les dimensions *Feature*, *Source*, *Time* et *Location*. Ceci est basé sur la supposition qu'un événement a nécessairement une caractéristique (par exemple, le niveau de lumière, la température, la présence de fumée, etc.), une source qui permet de détecter cette dernière (par exemple, un capteur de lumière, un capteur de température, un capteur de fumée, etc.), une heure et un lieu de son déclenchement.

Définition 13. Un **espace d'événement** $eSpace$ est défini comme un tuple de 6 attributs.

$$\langle id, Feature, Source, Time, Location, I \rangle$$

- **id** est l'identifiant unique de l'espace de l'événement.
- **Feature** est la dimension décrivant la caractéristique de l'événement.
- **Source** est la dimension décrivant le capteur pouvant détecter l'événement.
- **Time** est la dimension décrivant les données temporelles de l'événement.
- **Location** est la dimension décrivant les données spatiales de l'événement.
- **I** est l'ensemble des données d'instances de l'événement. L'élément **I** est vide lorsque l'événement est défini, puis des tuples de données sont insérés chaque fois que l'événement est déclenché. $I = \{ \langle id_1, Data_1 \rangle, \dots, \langle id_p, Data_p \rangle \}$, $id_{i \in 1, p}$ est l'identifiant de l'instance d'événement et $Data_{i \in 1, p}$ est l'ensemble des observations du capteur qui ont déclenché l'événement. $Data_{i \in 1, p} = \{ obs_1, \dots, obs_q \}$, $obs_{i \in 1, q} = \langle id_{obs}, sensor, time, location, value \rangle$. id_{obs} , $time$, $location$ et $value$ sont respectivement l'identifiant, le temps, le lieu et la valeur de l'observation. $sensor$ est le capteur qui a effectué l'observation. ■

Par exemple, si nous prenons l'événement gaspillage de lumière défini comme suit : $LightWastageEvent : \langle 1, 'LightWastage', eSpace_{LightWastage}, 'low' \rangle$

L'espace de cet événement peut être défini ainsi :

$eSpace_{LightWastage} : \langle 2, Feature_{LightWastage}, Source_{LightWastage}, Time_{LightWastage}, Location_{LightWastage}, I_{LightWastage} \rangle$, tel que :

- **Feature**_{LightWastage} : $\langle 3, Luminosity, \{Cost, Energy\}, \{hasFeature\}, \{Lumens > 30\} \rangle$
- **Source**_{LightWastage} : $\langle 4, LightSensor, \{LightingSystem, Lamp\}, \{senses, madeObservation\}, \{\} \rangle$
- **Time**_{LightWastage} : $\langle 5, TimeInterval, \{TemporalEntity, TimeInstant\}, \{inDateTime, inTemporalPosition\}, \{timestamp \textit{duration} \textit{1hour}, timestamp \textit{includedIn} [07:00:00pm, 07:00:00am]\} \rangle$
- **Location**_{LightWastage} : $\langle 6, Location, \{Coordinate, Position, Floor, Room, SpatialArea\}, \{hasLocation, hasAssignedOffice\}, \{\} \rangle$
- **I**_{LightWastage} : $\{ \langle 17, \{ \langle 36145, "lightSensor826", "28/08/13 19:00:00", "Office_413", 62 \rangle, \dots, \langle 36191, "lightSensor826", "28/08/13 20:00:00", "Office_413", 68 \rangle \} \rangle, \langle 18, \{ \langle 78569, lightSensor27, "03/10/13 22:48:00", "MeetingRoom_10", 62 \rangle, \dots, \langle 78569, "lightSensor826", "03/10/13 23:48:00", "MeetingRoom_10", 70 \rangle \} \rangle, \dots \}$

En plus des dimensions que nous avons détaillées un peu plus haut, l'eSpace d'un événement comporte également l'attribut I qui regroupe l'historique des données de déclenchement de l'événement. Nous pouvons voir dans l'exemple, les données de deux instances de déclenchement de l'événement gaspillage de lumière. La première instance a été détectée par le capteur "lightSensor826" dans l'emplacement "Office_413" entre 28/08/13 19:00:00 et 28/08/13 20:00:00, tandis que la deuxième a été détectée par le capteur "lightSensor27" dans l'emplacement "MeetingRoom_10" entre 28/08/13 22:48:00 et 28/08/13 23:48:00. Notez que ces instances dans I doivent absolument satisfaire les contraintes de déclenchement spécifiées dans les dimensions de l'eSpace ce qui est le cas ici puisque la luminosité est supérieure à 30 Lux durant le soir (cf. $\text{Feature}_{\text{LightWastage}}, \text{Time}_{\text{LightWastage}}$).

Après avoir présenté notre modèle pour la définition d'événements dans les environnements hybrides, dans la section suivante, nous détaillons notre modèle inspiré de l'approche 5W1H pour la définition des explications.

3.4.2 Un modèle pour la définition des explications des événements

Dans cette section, nous définissons l'explication d'un événement dans notre contexte et nous donnons quelques exemples. Comme nous l'avons expliqué précédemment (déf. 11, section 3.4.1), l'utilisateur peut attribuer trois niveaux d'urgence différents à l'événement (faible, moyen et élevé). Selon le niveau d'urgence, le système doit retourner une explication qui correspond au besoin de l'utilisateur en terme de contenu et de temps de traitement. Nous proposons donc trois niveaux d'explication différents, à savoir : *explication basique* e_b , *explication standard* e_s , et *explication complète* e_c qui correspondent respectivement à une urgence élevée, moyenne et faible d'une réponse à un événement. Nous rappelons ici que notre modèle s'inspire de l'approche 5W1H [74] et l'adapte au contexte de l'explication d'événements.

(i) **Explication basique de l'événement** : ce premier niveau d'explication contient les données de déclenchement de l'événement. Il vise à fournir à l'utilisateur les données nécessaires à une prise de décision rapide lorsque l'urgence de l'explication de l'événement est élevée (par exemple, un incendie). L'explication basique d'un événement est construite sur la base du traitement des données du déclenchement de ce dernier (c.-à-d. le tuple $\langle id_i, Data_i \rangle$ correspondant à l'événement déclenché dans la composante I de l'eSpace de l'événement, déf. 13).

Définition 14. *Une explication basique* e_b est basée sur l'approche 5W1H sans la composante *Why* puisque ce premier niveau d'explication est destiné à une situation urgente ($u = \text{high}$, déf. 11). e_b est formulée comme suit $e_b : \langle \text{What}, \text{Who}, \text{When}, \text{Where}, \text{How} \rangle$, où :

- **What** est la caractéristique de l'événement.
- **Who** est le capteur qui a déclenché l'événement.
- **When** est l'heure de déclenchement de l'événement.

- *Where* est le lieu de déclenchement de l'événement.
- *How* est l'ensemble des observations $Data_i$ qui ont déclenché l'événement. ■

Dans ce qui suit nous décrivons un exemple d'explication basique de l'événement gaspillage de lumière :

e_b : < **What** : Luminosity,
Who : LightSensor826,
When : 28/08/13 20:00:00,
Where : Office_413,
How : { < 36145, lightSensor826, 28/08/13 19:00:00, Office_413, 62 >, ...,
 < 36191, lightSensor826, 28/08/13 20:00:00, Office_413, 68 > } >

Comme vous pouvez le voir dans cet exemple, le champ *How* contient toutes les observations qui ont déclenché l'événement. Ces observations respectent les contraintes spécifiées dans la définition de l'événement gaspillage de lumière (section 3.4.1) : les niveaux de lumière sont supérieurs à 30 lux, ce qui correspond à la contrainte de la dimension $Feature_{LightWastage}$ ($Lumens > 30$) et les heures d'observation sont sur une plage horaire d'une heure après la plage horaire globale journalière, ce qui correspond aux contraintes de la dimension $Time_{LightWastage}$ (timestamp *duration* 1hour et timestamp *includedIn* [07:00:00pm, 07:00:00am]).

(ii) **Explication standard d'un événement** : ce niveau d'explication incorpore l'explication basique et ajoute un champ *Why* afin de correspondre à l'approche 5W1H. Lorsqu'un événement se produit, il y a une ou plusieurs entités dans l'environnement qui ont contribué à son déclenchement. Le champ *Why* va permettre de fournir une explication du déclenchement de l'événement (la cause du déclenchement). Afin qu'il soit lisible par l'utilisateur final, nous proposons de structurer lui-même ce champ *Why* en tuple 4W1H. En effet, pour fournir une explication de la cause du déclenchement d'un événement (le champ *Why*) on désire connaître le quoi lié à cette cause (c.-à-d. le *What* du *Why*), le qui lié à cette cause (c.-à-d. le *Who* du *Why*), le où lié à cette cause (c.-à-d. le *Where* du *Why*), le quand lié à cette cause (c.-à-d. le *When* du *Why*) et le comment lié à cette cause (c.-à-d. le *How* du *Why*). Ces tuples 4W1H sont exprimés sous la forme de concepts (par exemple les concepts *Employee*, *Office*, *LightingSystem*) sans spécifier exactement quelles instances (par exemple, *Roland_Perrin*, *Office413*, *Lamp45*) sont effectivement liées à l'événement qui a été déclenché. **Nous appelons le champ *Why* et les tuples contenus dans ce champ respectivement le champ Why_c et les tuples Why_c .** L'indice *c* fait référence ici au fait que ces éléments sont constitués de concepts (et non d'instances).

Notez que nous avons choisi de détailler les informations sur les causes potentielles du déclenchement de l'événement sous la forme de tuples 4W1H (les tuples Why_c) car cela permet d'exploiter pleinement l'approche 5W1H. De plus, structurer les raisons du déclenchement selon différentes facettes permet de faciliter

visuellement à l'utilisateur la compréhension. Enfin, cette structure nous permettra de calculer plus précisément les scores pour le classement des explications.

Définition 15. Une explication standard e_s d'un événement est composée de l'explication basique e_b de l'événement et d'un champ Why_c . Elle est formulée comme suit $e_s : \langle e_b, Why_c \rangle$. Ainsi, e_s constitue un tuple 5W1H, $e_s : \langle What, Who, When, Where, How, Why_c \rangle$, où :

- *What, Who, When, Where* et *How* sont les cinq champs qui constituent l'explication basique de l'événement e_b (déf. 14).
- Le champ Why_c est défini comme un ensemble de tuples 4W1H (**tuples Why_c**) associés à un score, $Why_c : \{t_{Why_{c_1}}, t_{Why_{c_2}}, \dots, t_{Why_{c_n}}\}$, $t_{Why_{c_i \in \{1, n\}}} : \langle c_{what}, c_{who}, c_{when}, c_{where}, c_{how}, score \rangle$ est un tuple Why_c , où :
 - **c_{what}** désigne l'ensemble des concepts décrivant les entités susceptibles de déclencher l'événement.
 - **c_{who}** est un concept identifié comme responsable du déclenchement de l'événement.
 - **c_{when}** désigne l'ensemble des concepts temporels qui relient c_{who} à l'événement.
 - **c_{where}** est l'ensemble des concepts spatiaux qui relient c_{who} à l'événement.
 - **c_{how}** est un ou plusieurs triplets : $\langle c_{who}, \textit{prédicat}, \textit{objet} \rangle$ qui déterminent la manière dont c_{who} a contribué au déclenchement de l'événement. Ici, le *prédicat* est la propriété reliant c_{who} à un *objet* et $\textit{objet} \in c_{what}$.
 - **score** représente le score de pertinence de l'explication qui est compris entre 0 et 1. ■

Par exemple, l'explication standard de l'événement gaspillage de lumière peut être représentée comme suit (les concepts sont indiqués en bleu) :

$$\begin{aligned}
 e_s : & \langle e_b, \\
 & \{ \langle c_{what} : \{\text{Lamp}\}, \\
 & c_{who} : \text{Employee}, \\
 & c_{when} : \{\text{TemporalEntity}\}, \\
 & c_{where} : \{\text{Office}, \text{MeetingRoom}, \text{BreakRoom}\} \\
 & c_{how} : \{ \langle \text{Employee}, \text{turnsOn}, \text{Lamp} \rangle, \langle \text{Employee}, \text{turnsOff}, \text{Lamp} \rangle \} \\
 & score : 0.85 \rangle, \dots \rangle
 \end{aligned}$$

e_b est l'explication basique de l'événement qui a été détaillée précédemment. Le champ Why_c comporte plusieurs *tuples Why_c* qui représentent les explications potentielles de la raison du déclenchement de l'événement gaspillage de lumière

associé à un score. L'explication présentée dans le champ Why_c de l'exemple suggère qu'un employé pourrait être responsable du déclenchement de l'événement (concept *Employee*, champ c_{who}), puisqu'il est celui qui manipule (allume ou éteint) les lampes installées dans l'environnement (concept *Lamp*, champs c_{what} et c_{how}). L'employé peut avoir plusieurs localisations spatio-temporelles qui le relie à l'événement (champs c_{when} et c_{where}). L'explication standard ne permet pas d'avoir le détail de cette localisation puisqu'elle ne traite que du niveau conceptuel. Néanmoins, un score de pertinence peut être calculé pour ce niveau d'explication. La construction de l'explication standard et la métrique de calcul de score seront détaillées dans la section 3.4.

(iii) **Explication complète de l'événement** : de la même façon, l'explication complète de l'événement e_c intègre l'explication standard e_s enrichie avec un nouveau champ Why contenant des détails supplémentaires. Ce champ Why est principalement basé sur des données de déclenchement de l'événement. Plus précisément, ce dernier complète le champ Why_c de l'explication standard (déf. 15) avec une couche supplémentaire d'informations. Cette couche est constituée des instances reliées à l'événement déclenché (par exemple, l'instance *Roland_Perrin* du concept *Employee* et l'instance *L015* du concept *Lamp*). Nous appelons cette version enrichie du champ Why_c , le champ **Why_i**. L'indice i fait référence ici au fait que le champ Why_i est constitué d'instances. Les tuples contenus dans le champs Why_i sont appelés **les tuples Why_i**.

Définition 16. Une explication complète e_c d'un événement est représentée comme suit : $e_c : \langle e_s, Why_i \rangle$. Le champ Why_i est défini comme un ensemble de tuples 4W1H (tuples Why_i) avec pour chacun un score de pertinence, $Why_i : \{t_{Why_{i_1}}, t_{Why_{i_2}}, \dots, t_{Why_{i_n}}\}, t_{Why_{i_{j \in \{1,n\}}}} : \langle i_{what}, i_{who}, i_{when}, i_{where}, i_{how}, score \rangle$ est un tuple Why_i , où les cinq premiers éléments sont des instances qui correspondent aux concepts de e_s et $score$ est une valeur comprise entre 0 et 1. ■

Par exemple, l'explication complète de l'événement gaspillage de lumière pourrait être représentée comme suit (les concepts et les instances sont représentés respectivement en bleu et en violet) :

$$\begin{aligned}
 e_c : & \langle e_s, \\
 & \langle i_{what} : \{\text{Lamp :L014, Lamp :L015}\}, \\
 & i_{who} : \text{Employee :Roland_Perrin}, \\
 & i_{when} : \text{TemporalEntity :28/08/13 18:32:50}, \\
 & i_{where} : \text{Office :Office413}, \\
 & i_{how} : \{\langle \text{Employee :Roland_Perrin, turnsOn, Lamp :L014} \rangle, \\
 & \langle \text{Employee :Roland_Perrin, turnsOn, Lamp :L015} \rangle\}, \\
 & score : 0.83 \rangle, \dots \rangle
 \end{aligned}$$

Comme vous le voyez, l'explication complète e_c ajoute une couche de détails à l'explication standard e_s présentée précédemment. Elle précise quel employé est responsable

du déclenchement de l'événement gaspillage de lumière (instance *Roland_Perrin*, champ i_{who}). Les lampes concernées (instances *L014* et *L015*, champs i_{what} et i_{how}) ainsi que l'heure et le lieu qui relie l'employé au déclenchement de l'événement (instances *28/08/13 18:32:50* et *Office413*, champs i_{when} et i_{where}) sont également précisés. Enfin, le score de pertinence de ce tuple Why_i est de 0.83. La construction de l'explication complète et la métrique de calcul de score seront détaillées respectivement dans les sections 3.5.3.2 et 3.5.3.3 .

En résumé, notre système ISEE vise à construire les explications potentielles du déclenchement des événements dans les environnements hybrides (section 3.4.2) à partir de la définition d'événements et des données de déclenchement (section 3.4.1). Pour structurer l'explication d'événements, le système ISEE propose trois niveaux d'explication différents (basique, standard et complet) qui s'inspirent du modèle 5W1H [72]. Nous synthétisons tous ces éléments dans notre modèle d'explication d'événements appelé modèle ISEE (figure 3.2). Il illustre le cycle de vie d'un événement (la définition de l'événement, son déclenchement et son explication) et détaille les connexions entre les différentes classes qui exploitent les définitions de nos deux modèles (sections 3.4.1 et 3.4.2).

Les algorithmes ISEE détaillés dans la section suivante visent à instancier les classes de la troisième couche de ce modèle (la couche explication de l'événement).

3.5 L'explication d'événements déclenchés dans un environnement hybride

Le système ISEE (figure 3.1) est basé sur l'hypothèse que l'analyse des interconnexions sémantiques entre les ontologies de domaine en s'appuyant sur les données de définition et de déclenchement d'événements permettra de construire des explications détaillées (hypothèse 2, section 1.2.3). Dans le chapitre 2, nous avons identifié les deux approches **d'alignement d'ontologies** [49] (section 2.4.1) et de **liaison de données** [49] (section 2.4.2) comme de bonnes pistes pour rapprocher les données de capteurs et les données du corpus documentaire. Par conséquent, dans cette section nous détaillons comment le système ISEE va tirer profit de ces deux approches pour construire les explications sur la base des 5W1H. Comme nous l'avons expliqué précédemment dans la section 5.2, le système ISEE est basé sur deux processus : un processus d'intégration des connaissances de domaine et le processus ISEE. Nous détaillons ces deux processus respectivement dans les sections 3.5.1 et 3.5.3. La section 3.5.2 présente la métrique d'évaluation d'alignement (*évaluation de l'alignement des ontologies*, figure 3.1) qui permet d'évaluer le résultat obtenu par la technique d'alignement (*alignement des ontologies*, figure 3.1) avant d'entamer le processus ISEE (*processus ISEE*, figure 3.1) .

Nous tenons à rappeler ici, que le processus d'intégration des connaissances de domaine ne constitue pas une contribution de cette thèse, nous y utilisons des outils de la littérature à des fins de préparation des données qui intègrent ensuite le processus

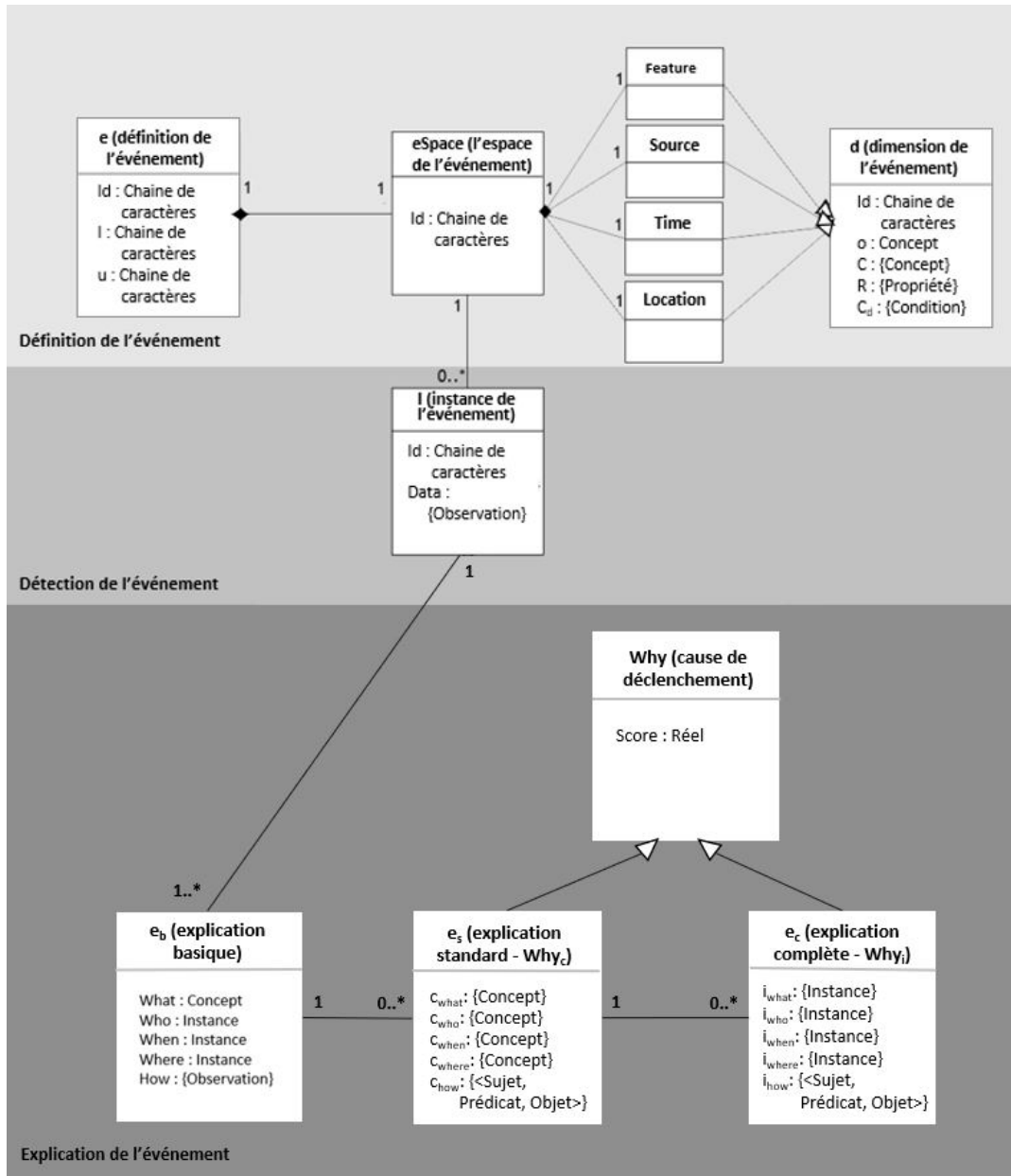


FIGURE 3.2 – Le modèle ISEE

ISEE.

3.5.1 Processus d'intégration des connaissances du domaine

Le processus d'intégration des connaissances de domaine a pour objectif d'intégrer toutes les informations issues de l'environnement (données de capteurs et du corpus documentaire) dans les ontologies de domaine sous forme d'instances. Ensuite, d'interconnecter ces ontologies, dans un premier temps, à l'aide de techniques d'alignement

classiques [49]. Il est donc constitué de deux étapes : (i) instanciation des ontologies de domaine et (ii) alignement de ces ontologies. Nous détaillons ces étapes respectivement dans les sections 3.5.1.1 et 3.5.1.2.

3.5.1.1 Instanciation des ontologies de domaine

Cette première étape a pour but d'alimenter les ontologies de domaine avec les données du réseau de capteurs et du corpus documentaire (*instanciation des ontologies*, figure 3.1). Pour ce faire, deux étapes sont nécessaires :

- Instanciation de l'ontologie HSSN-étendue (avec la notion d'événement, section 3.4.1) avec les données de capteurs.
- Instanciation des ontologies de domaine à partir du corpus documentaire.

Tout d'abord, pour traduire le flux de données de capteurs en instances d'ontologie plusieurs méthodes automatiques ou semi-automatiques peuvent être utilisées [24, 92, 104, 57, 140]. Les méthodes semi-automatiques nécessitent généralement que l'utilisateur précise des règles de mise en correspondance entre les données de capteurs et l'ontologie utilisée [20]. Selon la méthode de stockage et de gestion des données de capteurs utilisée (par exemple, SGBDR ou SGBDNoSQL, section 2.2.1), il existe de nombreux plug-in et plateformes gratuits, tels que par exemple RDBToOnto¹, morph² et morph-xR2RML³ qui permettent de traduire les données de capteurs (mais pas uniquement) en instances d'ontologie. La plateforme ISEE offre à l'utilisateur la flexibilité de choisir la méthode qui lui convient.

Pour analyser le corpus de documents et instancier les ontologies de domaine, plusieurs approches sont également proposées dans la littérature [38, 141, 38, 124, 151, 150]. Ces approches reposent sur plusieurs techniques (par exemple, NLP, apprentissage automatique, modèles statistiques, etc.) et diffèrent donc en terme de temps de traitement et de précision des résultats [38]. L'utilisateur de la plateforme ISEE a, ici également, le choix de la fonction d'instanciation qu'il souhaite utiliser. Ce processus est expliqué dans l'algorithme 1. L'algorithme parcourt le corpus de document et fait appel pour chaque ontologie à la fonction d'instanciation désignée en entrée f (algorithme 1, ligne 1-3). Celle-ci reçoit en entrée un document et l'ontologie de domaine que l'on souhaite instancier puis retourne ensuite l'ontologie mise à jour avec les instances nouvellement créées à partir du document (algorithme 1, ligne 3).

1. sourceforge.net/projects/rdbtoonto/

2. <https://github.com/jpcik/morph>

3. <https://github.com/frmichel/morph-xr2rml/>

Algorithme 1 Instanciation des ontologies de domaine avec le corpus de documents

Entrées : $D \leftarrow$ Corpus de documents hétérogènes

$O \leftarrow$ Ontologies de domaine

$f \leftarrow$ Fonction d'instanciation

```

1 pour chaque document de  $D$  faire
2   | pour chaque ontologie de  $O$  faire
3   |   |  $ontologie \leftarrow f(\text{document}, \text{ontologie})$ 
4   |   fin
5 fin

```

Noter ici que le processus d'instanciation des ontologies à partir du réseau de capteurs et à partir du corpus documentaire peuvent se faire en parallèle, ceci permet bien évidemment de gagner en terme de traitement.

Après avoir détaillé la première étape du processus d'intégration des connaissances de domaine, nous passons maintenant à la deuxième étape, celle de l'alignement des ontologies.

3.5.1.2 Alignement des ontologies de domaine

Cette étape vise à aligner les concepts des ontologies de domaine et de l'ontologie réseau de capteurs (*alignement des ontologies*, figure 3.1). Ces connexions sont construites hors-ligne indépendamment des données de déclenchement de l'événement. Elles constituent la base de notre système ISEE et permettent de naviguer de l'événement vers les concepts des ontologies de domaine. Nous détaillons ce processus dans l'algorithme 2.

Plusieurs techniques d'alignement peuvent être utilisées [137, 49]. Le plateforme ISEE offre à l'utilisateur la flexibilité de choisir l'outil d'alignement. Celui-ci est représenté dans l'algorithme par une fonction f . L'algorithme 2 parcourt toutes les ontologies et calcule la similarité sémantique entre chaque couple d'entités (concept ou propriété) à l'aide de la fonction f (Algorithme 2, ligne 6). Si le score de confiance de la correspondance est supérieur à un seuil, alors la correspondance est ajoutée à l'ensemble final de résultats (Algorithme 2, lignes 7-9).

Algorithme 2 Alignement des ontologies de domaine

entrées : $O \leftarrow$ Ontologies de réseau de capteurs et du corpus documentaire

$f \leftarrow$ Une fonction d'alignement

$seuil \leftarrow$ Le seuil d'alignement

sorties : Un ensemble de correspondances

```

1 Résultat  $\leftarrow$  {}
2 pour chaque  $o_i$  de  $O$  faire
3   pour chaque  $entité_p$  de  $o_i$  faire
4     pour chaque  $o_j$  de  $O$ , avec  $j > i$  faire
5       pour chaque  $entité_q$  de  $o_j$  faire
6          $\langle typeCorrespondance, score \rangle \leftarrow f(entité_p, entité_q)$ 
7         si  $score \geq seuil$  alors
8           Résultat  $\leftarrow$  Résultat  $\cup \langle entité_p, typeCorrespondance, entité_q, score \rangle$ 
9         fin
10      fin
11    fin
12  fin
13 fin
14 retourner Résultat

```

Par exemple, dans la figure 3.3, nous pouvons voir que la paire de concepts *Luminosity* de l'ontologie réseau de capteurs et *LightingSystem* de l'ontologie bâtiment est alignée avec une relation de spécialisation (\supset) qui a un score de confiance élevé (0,82). La correspondance $\langle Luminosity, \supset, LightingSystem, 0,82 \rangle$ est retournée dans les résultats de l'algorithme 2. Notez que toutes les correspondances ne sont pas représentées dans la figure 3.3, certaines ont été omises par souci de clarté.

Après avoir détaillé l'étape d'alignement des ontologies de domaine (*alignement des ontologies*, figure 3.1), dans la section suivante, nous présentons la métrique d'évaluation d'alignement (*évaluation de l'alignement des ontologies*, figure 3.1) qui permet à l'utilisateur d'évaluer le résultat retourné par l'outil d'alignement.

3.5.2 Évaluation de l'alignement des ontologies

La technique d'alignement (*alignement des ontologies* dans la figure 3.1, section 3.5.1.2) constitue le fondement de notre proposition et la base des étapes suivantes du système ISEE (étapes (1), (2) et (3), figure 3.1). Divers outils d'alignement existants peuvent être intégrés dans le système ISEE. L'exactitude des explications retournées par le système ISEE dépend directement de celle de l'alignement initial. Si le nombre ou la qualité des correspondances produites n'est pas satisfaisant, le système peut mal fonctionner ou retourner des résultats erronés. Par conséquent, nous proposons une métrique pour évaluer l'alignement. Cette métrique que nous appelons **la connectivité** mesure la capacité de l'outil d'alignement à connecter les ontologies de domaine. Elle est basée sur trois critères d'évaluation, à savoir la proportion des ontologies alignées (*Onto*), la proportion des concepts alignés (*Con*) et la moyenne des scores de confiance

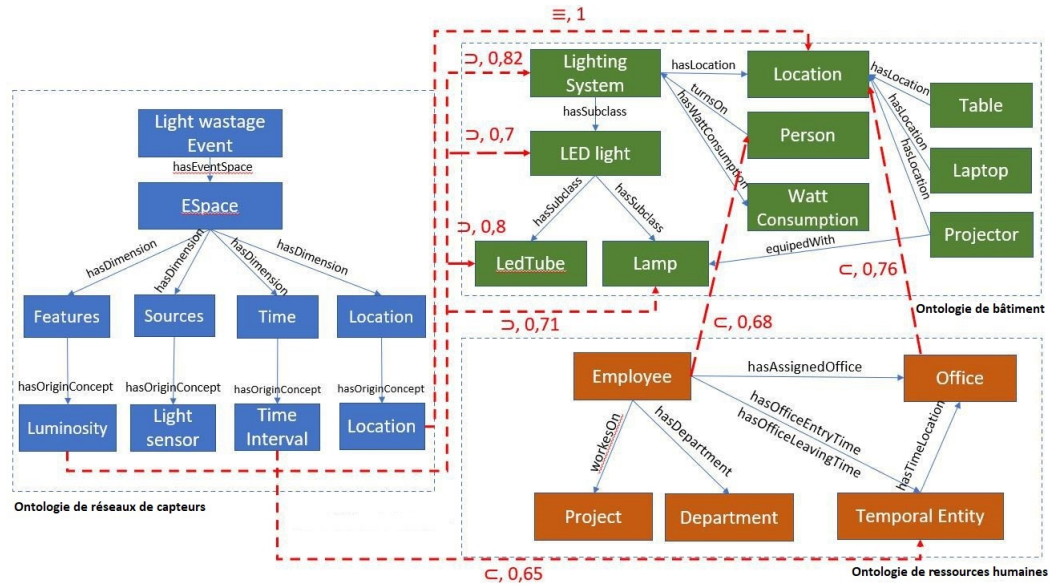


FIGURE 3.3 – Un exemple d’alignement d’ontologies

des correspondances (*Conf*). Le critère *Onto* évalue la capacité de la technique d’alignement à interconnecter le plus grand nombre d’ontologies. *Con* évalue la capacité de la technique d’alignement à interconnecter le plus grand nombre de concepts. *Conf* évalue la capacité de la technique d’alignement à construire des relations d’alignements avec des scores de confiance élevés. En effet, nous supposons que plus il y a de correspondances inter-ontologiques (*Onto*) et intra-ontologiques (*Con*) avec des scores de confiance élevés (*Conf*), plus il est possible d’accéder à un grand nombre d’informations pour expliquer un événement. Nous proposons donc la formule suivante :

$$\text{Connectivité}_{\text{alignement}} = \frac{1}{\alpha + \beta + \gamma} (\alpha \cdot \text{Onto} + \beta \cdot \text{Con} + \gamma \cdot \text{Conf})$$

- α , β et γ sont des coefficients de pondération.
- *Onto* désigne la proportion des ontologies alignées.

$$\text{Onto} = \frac{\text{Nombre d'ontologies alignées}}{\text{Nombre total d'ontologies}}.$$

- *Con* la proportion des concepts alignés.

$$\text{Con} = \frac{\text{Nombre de concepts alignés}}{\text{Nombre total de concepts}}.$$

- *Conf* désigne la moyenne des scores de confiance des correspondances produites.

$$\text{Conf} = \frac{1}{n} \sum_{i=1}^n c_i, \quad n \text{ est le nombre total de correspondances et } c_i \text{ le score de confiance de la } i\text{ème correspondance.}$$

Par exemple, si l’on fixe les valeurs de α , β et γ à 0.33, le score de connectivité de l’alignement présenté dans la figure 3.3 est égale à 0.74. *Onto* est égale à 1 puisque les trois ontologies sont alignées entre elles. *Con* est égale à 0.48 étant donné que 12 concepts sont alignés sur un total de 25 concepts. Enfin, *Conf* est égale à 0.76 la moyenne des scores de confiance des différents alignements. Notez ici que dans les expérimentations

menées pour évaluer le système ISEE (chapitre 4), nous avons fixé les coefficients de pondération à 1/3 respectivement. Cependant, la valeur de ces coefficients peut être modifiée par l'utilisateur du système. L'automatisation de la définition de ces coefficients sera traitée dans des travaux futurs.

Après avoir détaillé l'étape d'évaluation des alignements d'ontologies (*évaluation de l'alignement des ontologies*, figure 3.1), nous présentons dans la section suivante le processus ISEE qui constitue le cœur du système ISEE et englobe la deuxième contribution de ce manuscrit (section 1.3.2).

3.5.3 Le processus ISEE : Interconnexion et filtrage ciblés des ontologies de domaine

Dans cette section, nous présentons et détaillons la mise en oeuvre des trois étapes du processus ISEE (figure 3.1). Nous rappelons ici que ce processus est lancé automatiquement lorsqu'un événement est déclenché pour construire et retourner à l'utilisateur des pistes d'explications. **La première étape** du processus ISEE correspond au filtrage des concepts (figure 1.9, section 3.5.3.1). Elle s'appuie sur les alignements établis précédemment (*Interconnexion des concepts* dans la figure 1.9, section 3.5.1.2) pour construire un réseau d'interconnexions sémantiques sensibles au contexte entre l'événement et les concepts pertinents des différentes ontologies. Ces interconnexions sont exploitées pour construire l'explication standard e_s de l'événement (déf. 15). Ensuite, **la deuxième étape** effectue un deuxième filtrage au niveau des instances, pour ne garder que celles liées à l'événement déclenché. Sur la base de cet ensemble d'instances, un second réseau d'interconnexions sémantiques est construit (*filtrage des instances* dans la figure 1.9, section 3.5.3.2). Ces interconnexions sont exploitées pour construire l'explication complète e_c de l'événement (déf. 16). Enfin, **la dernière étape** analyse le graphe de connaissances ainsi construit, afin de classer les tuples Why_c (explication standard) ou Why_i (explication complète). Nous rappelons ici que les étapes (1) et (2) du processus ISEE constituent **la contribution 2.1** (section 1.3.2) de cette thèse, tandis que, l'étape (3) représente **la contribution 2.2** (section 1.3.2).

3.5.3.1 Interconnexions sémantiques au niveau des concepts

Les alignements classiques établis précédemment (section 3.5.1.2) permettent de rapprocher les ontologies à travers des connexions sémantiques d'équivalence, de généralisation et de spécialisation [49]. Dans notre contexte, ce type de connexion n'est pas suffisant pour fournir des réponses détaillées, nous avons besoin de connexions plus explicites pour guider le processus de recherche. Par conséquent, la première étape du processus ISEE se base sur l'alignement établi pour construire une deuxième couche de connexions sémantiques sensibles au contexte au niveau des concepts. Ces connexions sont exploitées ensuite pour construire l'explication standard e_s de l'événement (déf. 15). Concrètement, l'étape (1) du processus ISEE vise à établir des connexions sémantiques étiquetées $r_{C_{who}}$, $r_{C_{what}}$, $r_{C_{when}}$ et $r_{C_{where}}$ (r_c fait référence ici au fait que ces relations

sont des relations conceptuelles) entre l'événement déclenché et les concepts candidats à la réponse aux questions suivantes :

- **Connexions étiquetées rc_{who}** : quel concept désigne l'entité responsable du déclenchement de l'événement ?
- **Connexions étiquetées rc_{what}** : quel concept explique la relation de cette entité avec le déclenchement de l'événement ?
- **Connexions étiquetées rc_{where}** : quel concept est susceptible de la relier au lieu du déclenchement de l'événement ?
- **Connexions étiquetées rc_{when}** : quel concept est susceptible de la relier à l'heure du déclenchement de l'événement ?

Par exemple, si nous prenons l'événement gaspillage de lumière défini précédemment (section 3.4), les quatre concepts *Employee*, *Lamp*, *Temporal Entity* et *Office* représentent respectivement des candidats à la réponse aux quatre questions susmentionnées (cf. exemple de l'explication standard de l'évènement gaspillage de lumière, section 3.4.2). Le concept *Employee* représente l'entité responsable du déclenchement de l'évènement gaspillage de lumière. La relation entre le concept *Employee* et l'évènement gaspillage de lumière est expliquée par le concept *Lamp* (l'employé est celui qui manipule la lampe qui est une source de lumière pouvant déclencher l'évènement). Les concepts *Temporal Entity* et *Office* sont susceptibles respectivement d'expliquer la relation temporelle et spatiale entre le concept *Employee* et l'évènement gaspillage de lumière.

Le graphe de connaissances produit dans cette étape est utilisé pour identifier les causes potentielles du déclenchement de l'évènement et ainsi **construire les tuples Why_c de l'explication standard** (*explication standard*, figure 3.2) puisque nous ne considérons ici que le niveau conceptuel. Dans ce qui suit, nous expliquons d'abord la motivation et le principe général de la construction des quatre types d'interconnexions rc_{who} , rc_{what} , rc_{when} et rc_{where} . Ensuite, nous détaillons la démarche proposée et nous présentons quelques exemples d'explications standards.

>>> Motivation et principe général

Afin de construire les interconnexions étiquetées rc_{who} , rc_{what} , rc_{when} et rc_{where} , nous nous appuyons sur les alignements produits à l'étape 2 du processus d'intégration des connaissances de domaine (*processus d'intégration des connaissances de domaine* dans la figure 3.1, section 3.5.1.2) et les concepts représentant chacune des dimensions *Feature*, *Source*, *Time* et *Location* de l'évènement (déf. 13).

- **Les connexions étiquetées rc_{who} et rc_{what}** : les connexions étiquetées rc_{who} relient l'évènement aux concepts susceptibles d'être responsables de son déclenchement, tandis que les connexions étiquetées rc_{what} relient l'évènement aux concepts qui expliquent pourquoi ils ont été identifiés comme responsables. Pour établir les connexions étiquetées rc_{who} , nous nous basons sur deux hypothèses :

- **Hypothèse (1)** : pour qu'un concept soit responsable du déclenchement d'un évènement dans un environnement, il doit faire partie de ce dernier. En d'autres termes, le concept doit représenter une entité physique appartenant à l'environnement (par exemple les concepts *Employee* et *Projector*, déf. 9)
- **Hypothèse (2)** : la deuxième hypothèse consiste à dire que le concept responsable du déclenchement d'un événement doit être relié aux caractéristiques de ce dernier, c'est-à-dire à sa dimension *Feature* (par exemple, pour qu'on puisse dire le concept *Employee* est un responsable potentiel du déclenchement de l'évènement gaspillage de lumière celui-ci doit avoir une relation qui le lie au concept *Luminosity*, l'origine de la dimension *Feature* de cet évènement, déf. 13).

Si le concept valide ces deux hypothèses alors une connexion étiquetée rc_{who} est établie entre l'évènement et ce dernier. De plus, des connexions étiquetées rc_{what} sont également établies entre ce concept et ceux qui ont permis de le relier à l'évènement (par exemple, le concept *Lamp* est celui qui nous a permis de dire que le concept *Employee* est un responsable potentiel du déclenchement de l'évènement puisque que ce dernier est celui qui manipule les lampes pouvant déclencher l'évènement gaspillage de lumière).

- **Les connexions étiquetées rc_{when}** : ces connexions relient l'évènement aux concepts qui expliquent la relation temporelle entre l'évènement et le concept responsable de son déclenchement, par exemple, le concept *TemporalEntity* explique la relation entre le concept *Employee* et l'évènement gaspillage de lumière (cf. exemple de l'explication standard de l'évènement gaspillage de lumière, section 3.4.2). Pour établir ces connexions, nous nous appuyons tout naturellement sur la dimension *Time* de l'évènement (déf. 13) et plus précisément sur les concepts alignés au concept origine de cette dimension (attribut *o*, déf. 12).
- **Les connexions étiquetées rc_{where}** : ces connexions relient l'évènement aux concepts qui expliquent la relation spatiale entre l'évènement et l'entité responsable de son déclenchement, par exemple, le concept *Office* explique la relation spatiale entre le concept *Employee* et l'évènement gaspillage de lumière (cf. exemple de l'explication standard de l'évènement gaspillage de lumière, section 3.4.2). De la même façon que les connexions rc_{when} , pour établir les connexions rc_{where} , nous nous appuyons sur la dimension *Location* de l'évènement (déf. 13) et plus précisément sur les concepts alignés au concept origine de cette dimension (attribut *o*, déf. 12).

Maintenant que nous avons présenté la motivation et le principe général de la construction des quatre types d'interconnexions, nous détaillons dans ce qui suit notre démarche.

>>> Démarche proposée

L'algorithme 3 se décompose en deux parties principales : la première partie a pour but d'établir les interconnexions étiquetées rc_{what} , rc_{when} et rc_{where} en suivant la même démarche (lignes 2-7, algorithme 3). La deuxième partie établit les interconnexions étiquetées rc_{who} (lignes 8-21, algorithme 3). Nous détaillons dans ce qui suit chacune de ces deux parties.

— **Construction d'interconnexions étiquetées rc_{what} , rc_{when} et rc_{where}**

Pour établir **les connexions rc_{what}** qui relient l'évènement aux concepts expliquant pourquoi une entité est responsable de son déclenchement, nous exploitons les relations d'alignement qui relient **le concept origine de la dimensions *Feature*** (attribut *o*, déf. 12) avec les concepts du graphe de connaissance. Nous parcourons itérativement les concepts alignés avec le concept origine de la dimension *Feature* puis nous établissons une relation étiquetée rc_{what} entre ces concepts et l'évènement (ligne 4-6, algorithme 3, hypothèse (2)). Par exemple, dans la figure 3.4, la dimension *Feature* de l'évènement *light wastage* (gaspillage de lumière) est représentée par le concept origine *Luminosity*. En outre, les concepts *Luminosity* et *Lighting System* sont alignés par une relation de spécialisation (représentée en rouge). Ainsi, une connexion sémantique étiquetée rc_{what} est établie entre l'évènement *light wastage* et le concept *Lighting System*. Nous procédons exactement de la même façon pour les deux dimensions *Time* et *Location*.

— **Construction d'interconnexions étiquetées rc_{who}**

Pour construire les connexions sémantiques étiquetées rc_{who} , qui relient l'évènement aux concepts responsables de son déclenchement, nous utilisons les concepts précédemment reliés à l'évènement par les connexions sémantiques étiquetées rc_{what} . Nous parcourons itérativement leurs voisins directs puis nous sélectionnons uniquement ceux qui représentent une entité de l'environnement (ligne 10-15, algorithme 3, hypothèse (1)). Une connexion sémantique étiquetée rc_{who} est alors établie entre l'évènement et ces concepts que nous appelons *concept_{who}* (ligne 14, algorithme 3).

Algorithme 3 Connexions rc_{what} , rc_{who} , rc_{when} et rc_{where} des concepts

entrées : $O \leftarrow$ Ontologies de réseaux de capteurs et du corpus documentaire
RésultatAlignements \leftarrow Alignements des ontologies de réseaux de capteurs et du corpus documentaires produits par l'algorithme 2
événement \leftarrow Définition de l'événement
Entités \leftarrow Ensemble des concepts qui représentent les entités de l'environnement

sorties : Un ensemble de triplets \langle événement, prédicat, concept \rangle avec rc_{what} , rc_{who} , rc_{when} , et rc_{where} comme prédicats.

```

1 Résultat  $\leftarrow$  {}
2  $o_{Feature} \leftarrow$  événement.eSpace.Feature.o
3 ConceptsWhat  $\leftarrow$  ObtenirLesConceptsAlignés( $o_{Feature}$ , RésultatAlignements)
4 pour chaque  $concept_{what}$  de ConceptsWhat faire
5   | Résultat  $\leftarrow$  Résultat  $\cup$   $\langle$ événement.id,  $rc_{what}$ ,  $concept_{what}$   $\rangle$ 
6 fin
   /* faire la même chose pour les dimensions Time et Location pour construire
   respectivement les triplets avec les prédicats When et Where */
7 [...]
8  $o_{Source} \leftarrow$  événement.eSpace.Source.o
9 ConceptsWho  $\leftarrow$  {}
10 pour chaque  $concept_{what}$  de ConceptsWhat faire
11   | ConceptsWho  $\leftarrow$  ObtenirLesConceptsVoisins( $concept_{what}$ )
12   | pour chaque  $concept_{who}$  de ConceptsWho faire
13     | si  $concept_{who} \in$  Entités alors
14       | Résultat  $\leftarrow$  Résultat  $\cup$   $\langle$ événement.id,  $rc_{who}$ ,  $concept_{who}$   $\rangle$ 
15       | ConceptsAlignésAuConceptsWho  $\leftarrow$  ObtenirLesConceptsAlignés( $concept_{who}$ , RésultatAlignements)
16       | pour chaque  $c$  de ConceptsAlignésAuConceptsWho faire
17         | Résultat  $\leftarrow$  Résultat  $\cup$   $\langle$ événement.id,  $rc_{who}$ ,  $c$   $\rangle$ 
18         | fin
19       | fin
20     | fin
21   | fin
22 retourner Résultat

```

Ensuite, l'algorithme exploite les résultats d'alignement pour trouver les potentiels concepts responsables du déclenchement de l'événement qui n'ont pas pu être identifiés parce qu'ils n'ont pas une relation directe avec la dimension *Feature*. Pour ce faire, tous les concepts qui ont une relation d'alignement avec les $concept_{who}$ sont également reliés à l'événement avec une connexion étiquetée rc_{who} (ligne 15-18, algorithme 3). Par exemple, le concept *Lighting System* (relié précédemment à l'événement avec une connexion étiquetée rc_{what}) a trois voisins : *Location*, *Person* et *WattConsumption*. Ainsi, une connexion sémantique étiquetée rc_{who} est alors établie entre l'événement *light wastage* et le concept *Person* puisque celui-ci représente une entité de l'environnement. De plus, comme les concepts *Person* et *Employee* ont été précédemment liés par une relation de spécialisation, une connexion sémantique étiquetée rc_{who} est également établie entre l'événement *light wastage* et le concept *Employee*.

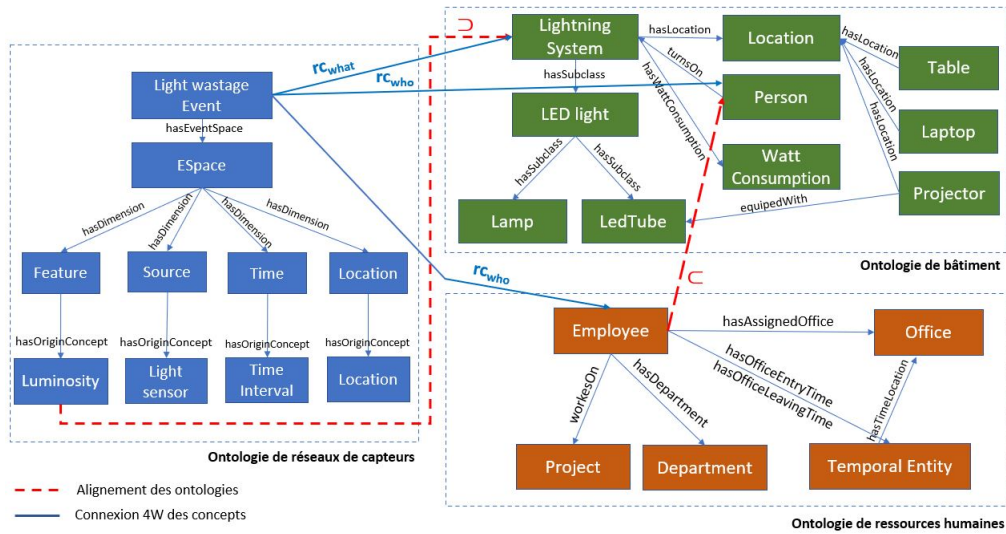


FIGURE 3.4 – Un exemple de connexions rc_{what} et rc_{who} de concepts

Pour des raisons de simplicité, nous ne détaillons pas ici toutes les interconnexions sémantiques étiquetées rc_{what} et rc_{who} .

Après avoir détaillé notre démarche de construction d'interconnexions étiquetées rc_{what} , rc_{who} , rc_{when} et rc_{where} , nous expliquons à présent comment, à partir de ces interconnexions, nous allons construire l'explication standard de l'événement (déf. 15).

»»» Construction des explications standards des événements

Comme nous l'avons indiqué précédemment (section 3.4.2), l'explication standard (déf. 15, e_s dans la figure 3.2) de l'événement est constituée de l'explication basique (déf. 14, e_b dans la figure 3.2) qui contient les données du déclenchement de l'événement et d'un champ Why_c . Le champ Why_c détaille les candidats d'explication de l'événement au niveau conceptuel. **Il est constitué d'un ensemble de tuples Why_c associés à un score** (déf. 15). **Chaque tuple Why_c détaille une piste d'explication.** Dans cette section, nous expliquons comment à partir des interconnexions sémantiques rc_{what} , rc_{who} , rc_{when} et rc_{where} , nous allons pouvoir renseigner les tuples Why_c , ainsi construire l'explication standard e_s de l'événement (e_s , figure 3.2). Ce processus est détaillé dans l'algorithme 4.

— Identification des attributs c_{what} , c_{who} et c_{how} du tuple Why_c

Dans l'étape précédente nous avons établi des interconnexions sémantiques étiquetées rc_{what} , rc_{who} , rc_{when} et rc_{where} . Ces interconnexions relient l'événement respectivement aux candidats c_{what} , c_{who} , c_{when} et c_{where} des tuples Why_c de l'explication standard (déf. 15). Cependant, ces candidats sont actuellement déconnectés les uns des autres et ne constituent pas encore des tuples (tuples Why_c , déf. 15). Le premier objectif de cette étape est

d'identifier les candidats c_{what} et c_{who} qui sont reliés. Pour ce faire, l'algorithme analyse tout d'abord itérativement les voisins de chaque concept c_{who} et sélectionne ceux qui représentent des concepts c_{what} (algorithme 4, lignes 7-8, 11-12). De plus, l'algorithme exploite également les relations d'alignement pour identifier de nouveaux candidats. Tout concept qui a une relation d'alignement avec les concepts c_{what} est également identifié comme nouveau concept c_{what} (algorithme 4, lignes 9-10, 11-12).

Algorithme 4 Construction des tuples Why_c

entrées : $O \leftarrow$ Ontologies de réseaux de capteurs et du corpus documentaire
 RésultatAlignements \leftarrow Alignements des ontologies de réseaux de capteurs et du corpus documentaires produits par l'algorithme 2
 Résultat_étape1 \leftarrow Les triples étiquetés What, Who, Where et When fournis par l'algorithme

sorties : Un ensemble de tuple Why_c

```

1 Résultat  $\leftarrow$  {}
2 ConceptsWhat  $\leftarrow$  ObtenirLesConceptsWhat(Résultat_step1)
  /* faire la même chose pour les concepts Who, When et Where */
3 pour chaque conceptwho de ConceptsWho faire
4   tWhyc  $\leftarrow$  TupleWhyc()
5   tWhyc.cwho  $\leftarrow$  conceptwho
6   tWhyc.cwhat  $\leftarrow$  {}
  /* faire la même chose pour tWhyc.cwhen et tWhyc.cwhere */
7   pour chaque conceptwhat de ConceptsWhat faire
8     VoisinsDirectConceptWhat  $\leftarrow$  ObtenirLesConceptsVoisins(conceptwhat)
9     ConceptsAlignésAuConceptsWhat  $\leftarrow$  ObtenirLesConceptsAlignés(conceptwhat, RésultatAlignements)
10    VoisinsIndirectConceptWhat  $\leftarrow$  ObtenirLesConceptsVoisins(ConceptsAlignésAuConceptsWhat)
11    si conceptwho  $\in$  VoisinsDirectConceptWhat ou conceptwho  $\in$  VoisinsIndirectConceptWhat alors
12      tWhyc.cwhat  $\leftarrow$  IWhy.cwhat  $\cup$  conceptwhat
13      PropriétéReliantWhatEtWho  $\leftarrow$  ObtenirPropriétéReliant(conceptwhat, conceptwho)
14      tWhyc.CandidatHow  $\leftarrow$  tWhyc.CandidatHow  $\cup$  < conceptwhat, PropriétéReliantWhatEtWho, conceptwho >
15    fin
16  fin
17  VoisinsConceptWho  $\leftarrow$  ObtenirLesConceptsVoisins(conceptwho)
18  pour chaque conceptwhen de ConceptsWhen faire
19    si conceptwhen  $\in$  VoisinsConceptWho alors
20      tWhyc.cwhen  $\leftarrow$  tWhyc.cwhen  $\cup$  conceptwhen
21    fin
22  fin
  /* faire la même chose pour les concepts Where */
23 Résultat  $\leftarrow$  Résultat  $\cup$  tWhyc
24 fin
25 retourner Résultat
    
```

Par exemple, dans la figure 3.4, le concept c_{who} *Person* de l'ontologie bâtiment est relié directement au concept c_{what} *Lighting System*. Par conséquent, les concepts *Person* et *Lighting System* sont agrégés dans un tuple Why_c respectivement comme attributs c_{who} et c_{what} . De plus, le concept *Employee*

de l'ontologie ressources humaines est aligné avec le concept *Person* de l'ontologie bâtiment, par conséquent, un autre tuple Why_c est construit avec comme attributs c_{who} et c_{what} les concepts *Employee* et *Lighting System*. Afin d'ajouter aux tuples Why_c ainsi construits l'attribut c_{how} , nous exploitons les propriétés reliant les attributs c_{who} et c_{what} . Par exemple, les attributs *Person* et *Lighting System* du tuple Why_c construit précédemment sont reliés par les propriétés *turnsOn* et *turnsOff*. Par conséquent, les triplets $\langle Person, turnsOn, Lighting System \rangle$ et $\langle Person, turnsOff, Lighting System \rangle$ sont utilisés pour renseigner l'attribut c_{how} du tuple Why_c (algorithme 4, lignes 13-14). Dans le deuxième exemple constitué des attributs *Employee* et *Lighting System*, les tuples $\langle Employee, turnsOn, Lighting System \rangle$ et $\langle Employee, turnsOff, Lighting System \rangle$ sont utilisés pour renseigner l'attribut c_{how} en s'appuyant toujours sur la relation d'alignement. Nous rappelons ici que l'attribut c_{how} a pour rôle d'expliquer à l'utilisateur la relation entre l'entité responsable du déclenchement de l'événement (par exemple le concept *Employee*) et l'élément qui a permis de l'identifier comme responsable (par exemple le concept *Lighting System*).

— **Identification des attributs c_{when} et c_{where} du tuple Why_c**

Enfin, pour renseigner les attributs c_{when} et c_{where} du tuple Why_c , l'algorithme cherche les candidats c_{when} et c_{where} reliés directement aux concepts c_{who} (algorithme 4, lignes 17-23). Par exemple, *TemporalEntity* et *Office* sont respectivement des candidats c_{when} et c_{where} reliés directement au concept *Employee* (concept c_{who}). Par conséquent, ces deux concepts sont utilisés respectivement pour renseigner les attributs c_{when} et c_{where} du tuple Why_c . La figure 3.5 montre l'explication standard construite à partir des deux exemples cités précédemment. Comme vous pouvez le constater le champ Why_c contient deux tuples Why_c . Certains attributs du tuple Why_{c2} ne sont pas renseignés étant donné que nous n'avons pas pu trouver des informations dans les ontologies qui nous permettent de les renseigner. Enfin, la métrique de calcul de score sera détaillée dans la section 3.5.3.3.

3.5.3.2 Interconnexions sémantiques au niveau des instances

Le déclenchement d'un événement fournit de nombreuses informations intéressantes, notamment le contexte spatio-temporel (par exemple, l'heure et le lieu du déclenchement de l'événement gaspillage de lumière, la liste des équipements pouvant produire de la lumière dans le bureau 413, le rapport d'accès au badge des bureaux, etc.). Jusqu'à présent, nous n'avons pas utilisé ces informations pour construire les explications car la première étape du processus ISEE (section 3.5.3.1) était dédiée aux événements ayant un niveau d'urgence moyen et ne traitait donc que du niveau conceptuel. La deuxième étape du processus ISEE quant à elle, est consacrée aux événements dont le niveau d'urgence est bas. Nous pouvons donc aller plus loin et analyser les instances du graphe sémantique. Cette deuxième étape du processus ISEE se base sur les

What	Luminosity				
Who	LightSensor44				
When	28/08/13 20:00:00				
Where	Office 413				
How	< 36145, lightSensor44, 28/08/13 19:00:00, Office_413, 62 >, ..., < 36191, lightSensor44, 28/08/13 20:00:00, Office_413, 68 >				
Why _c	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 50%; padding: 5px;"> <pre> c_{What} : Lamp c_{Who} : Employee c_{When} : TemporalEntity c_{Where} : Office c_{How} : < Employee, turnsOn, Lamp > < Employee, turnsOff, Lamp > Score = 0,88 </pre> <p style="text-align: center;">Tuple Why_{c1}</p> </td> <td style="width: 50%; padding: 5px;"> <pre> c_{What} : Lamp c_{Who} : Person c_{When} : - c_{Where} : - c_{How} : < Person, turnsOn, Lamp > < Person, turnsOff, Lamp > Score = 0,65 </pre> <p style="text-align: center;">Tuple Why_{c2}</p> </td> </tr> <tr> <td colspan="2" style="text-align: center;">Tuples Why_c</td> </tr> </table>	<pre> c_{What} : Lamp c_{Who} : Employee c_{When} : TemporalEntity c_{Where} : Office c_{How} : < Employee, turnsOn, Lamp > < Employee, turnsOff, Lamp > Score = 0,88 </pre> <p style="text-align: center;">Tuple Why_{c1}</p>	<pre> c_{What} : Lamp c_{Who} : Person c_{When} : - c_{Where} : - c_{How} : < Person, turnsOn, Lamp > < Person, turnsOff, Lamp > Score = 0,65 </pre> <p style="text-align: center;">Tuple Why_{c2}</p>	Tuples Why _c	
<pre> c_{What} : Lamp c_{Who} : Employee c_{When} : TemporalEntity c_{Where} : Office c_{How} : < Employee, turnsOn, Lamp > < Employee, turnsOff, Lamp > Score = 0,88 </pre> <p style="text-align: center;">Tuple Why_{c1}</p>	<pre> c_{What} : Lamp c_{Who} : Person c_{When} : - c_{Where} : - c_{How} : < Person, turnsOn, Lamp > < Person, turnsOff, Lamp > Score = 0,65 </pre> <p style="text-align: center;">Tuple Why_{c2}</p>				
Tuples Why _c					

FIGURE 3.5 – Un exemple d’explication standard de l’événement gaspillage de lumière

données issues du déclenchement de l’événement et sur le graphe sémantique construit par l’algorithme 3 (connexions rc_{what} , rc_{who} , rc_{when} et rc_{where} des concepts) pour établir de nouvelles connexions au niveau des instances. Ces connexions seront ensuite analysées pour construire l’explication complète e_c de l’événement (e_c , figure 3.2).

Concrètement, l’étape (2) s’appuie sur les données issues du déclenchement de l’événement et sur le résultat de l’algorithme 3 pour établir des connexions sémantiques étiquetées ri_{who} , ri_{what} , ri_{when} et ri_{where} (ri fait référence ici au fait que ces relations sont des relations aux niveau des instances) entre l’événement et les instances candidates à la réponse aux questions suivantes :

- **Connexions étiquetées ri_{who}** : quelle instance désigne l’entité responsable du déclenchement de l’événement ?
- **Connexions étiquetées ri_{what}** : quelle instance explique la relation de cette entité avec le déclenchement de l’événement ?
- **Connexions étiquetées ri_{when}** : quelle instance est susceptible de la relier à l’heure du déclenchement de l’événement ?
- **Connexions étiquetées ri_{where}** : quelle instance est susceptible de la relier au lieu du déclenchement de l’événement ?

Si nous prenons l’exemple de l’événement gaspillage de lumière déclenché dans le bureau 413 à 20:00:00 le 28/08/13 (cf. exemple de l’explication complète de l’évènement gaspillage de lumière, section 3.4.2), les quatre instances *Roland Perrin* (instance du concept *Employee*), *L014* (instance du concept *Lamp*), *28/08/13 18:32:50* (instance du concept *TemporalEntity*) et *Office413* (instance du concept *Office*) représentent respectivement des candidats à la réponse aux quatre questions susmentionnées. L’instance *Roland Perrin* représente l’entité responsable du déclenchement de l’événement

gaspillage de lumière. La relation entre l'instance *Roland_Perrin* et l'évènement gaspillage de lumière est expliquée par l'instance *L014* (l'employé *Roland_Perrin* est celui qui a laissé la lampe allumée, la lampe étant une source de lumière). Enfin, les instances *28/08/13 18:32:50* et *Office413* relient *Roland_Perrin* à l'évènement déclenché (*Roland_Perrin* a quitté le bureau 413 à 18:32:50 le 28/08/13 peu de temps avant le déclenchement de l'évènement gaspillage de lumière).

Le graphe de connaissances produit à cette étape est utilisé pour construire l'explication complète de l'évènement (*explication complète*, figure 3.2). Dans ce qui suit, nous expliquons d'abord la motivation et le principe général de la construction des quatre types d'interconnexions ri_{who} , ri_{what} , ri_{when} et ri_{where} . Ensuite, nous détaillons la démarche proposée et nous présentons des exemples.

>>> **Motivation et principe général**

Afin de construire les interconnexions étiquetées ri_{who} , ri_{what} , ri_{when} et ri_{where} au niveau des instances, nous nous appuyons sur les données issues du déclenchement de l'évènement (le capteur qui a déclenché l'évènement, les données capturées par ce dernier, l'heure et le lieu du déclenchement de l'évènement, attribut *I* dans la déf. 13). Nous rappelons ici que pour établir ces interconnexions, nous ne nous intéressons pas à l'ensemble du graphe sémantique mais plutôt aux instances des concepts précédemment liés à l'évènement avec les connexions étiquetées rc_{who} , rc_{what} , rc_{when} et rc_{where} (résultat de l'algorithme 3).

- **Les connexions étiquetées ri_{who} , ri_{when} et ri_{where}** : les connexions étiquetées ri_{who} relient l'évènement aux instances susceptibles d'être responsable de son déclenchement, ces instances étant des entités de l'environnement (hypothèse (1), section 3.5.3.1). Les connexions étiquetées ri_{when} et ri_{where} relient l'évènement respectivement aux instances qui expliquent la relation temporelle et spatiale entre l'évènement et les entités responsables de son déclenchement. Pour construire **les connexions étiquetées ri_{who}** nous faisons l'hypothèse suivante :

- **Hypothèse (3)** : pour qu'une instance soit reliée au déclenchement d'un évènement, celle-ci doit avoir une localisation spatio-temporelle proche de celle du déclenchement de cet évènement.

Cette hypothèse repose sur l'idée que plus une instance est proche du lieu et de l'heure du déclenchement d'un évènement, plus elle est susceptible d'y être impliquée. Par conséquent, si une instance d'un concept précédemment relié à l'évènement avec une connexion rc_{who} valide cette hypothèse, celle-ci est considérée comme potentiel responsable du déclenchement de l'évènement. Une nouvelle connexion étiquetée ri_{who} est alors établie entre l'évènement et cette instance (par exemple, l'instance *Roland_Perrin* du concept *Employee* dans le cas de l'évènement gaspillage de lumière). De plus, des **connexions étiquetées ri_{when} et ri_{where}** sont établies respectivement entre l'évènement et les instances qui ont permises de la relier à la posi-

tion spatiale (par exemple, l'instance *Office413*) et temporelle (par exemple, l'instance *28/08/13 18:32:50*) de l'événement.

- **Les connexions étiquetées ri_{what}** : les connexions étiquetées ri_{what} relient l'évènement aux instances qui expliquent les raisons pour lesquelles certaines entités ont été identifiées comme responsables du déclenchement de cet événement, (par exemple, l'instance *L014* du concept *Lamp* explique pourquoi *Roland_Perrin* a été identifié comme potentiel responsable du déclenchement de l'événement gaspillage de lumière, étant donné que celui-ci a oublié cette lampe allumée, la lampe étant une source de lumière). Pour établir **les connexions étiquetées ri_{what}** , nous nous appuyons bien évidemment sur les concepts précédemment reliés à l'événement avec des connexions rc_{what} . Nous faisons l'hypothèse suivante :

- **Hypothèse (4)** : pour qu'une instance puisse expliquer la relation d'une entité avec le déclenchement de l'événement, celle-ci doit valider l'hypothèse (3) et doit avoir une relation sémantique avec l'entité en question.

Si une instance d'un concept précédemment relié à l'événement avec une connexion rc_{what} valide cette hypothèse, une nouvelle relation ri_{what} est établie entre l'événement et cette instance. Par exemple, l'instance *L014* du concept *Lamp* est reliée à l'entité *Roland_Perrin* (identifiée plus haut comme potentiel responsable du déclenchement de l'événement gaspillage de lumière) avec deux relations sémantiques *turnsOn* et *turnsOff*. De plus, l'instance *L014* a comme localisation *Office413* le lieu du déclenchement de l'événement gaspillage de lumière. Par conséquent, une nouvelle connexion étiquetée ri_{what} est établie entre l'instance *L014* et l'événement gaspillage de lumière. L'instance *L014* explique la relation entre l'entité *Roland_Perrin* et l'événement gaspillage de lumière (l'employé Roland Perrin est peut-être celui qui a oublié d'éteindre la lumière dans le bureau 413).

Maintenant que nous avons présenté la motivation et le principe général de la construction des quatre types d'interconnexions, nous détaillons à présent notre démarche.

>>> Démarche proposée

L'algorithme 5 se décompose en deux parties principales : la première partie a pour but d'établir les interconnexions étiquetées ri_{who} , ri_{when} et ri_{where} (lignes 2-14, algorithme 5). La deuxième partie établit les interconnexions étiquetées ri_{what} (lignes 15-25, algorithme 5). Nous détaillons dans ce qui suit chacune de ces deux parties.

- **Construction d'interconnexions étiquetées ri_{who} , ri_{when} et ri_{where}**
 Pour construire les connexions ri_{who} , ri_{when} et ri_{where} , les instances des concepts précédemment interconnectés à l'événement par le biais des connexions sémantiques rc_{who} sont analysées afin de sélectionner celles dont le

lieu et l'heure sont proches du déclenchement de l'événement (lignes 2-14, algorithme 5). Pour ce faire, les techniques de liaison de données sont utilisées pour comparer l'historique des positions des instances des concepts rc_{who} au lieu et l'heure du déclenchement de l'événement (ligne 7, algorithme 5). Si la fonction de liaison de données identifie une position d'une instance comme similaire à celle du déclenchement de l'événement (ligne 8, algorithme 5), l'instance est connectée à l'événement avec une nouvelle relation étiquetée ri_{who} (ligne 9, algorithme 5). Les instances qui ont permises d'avoir la position spatio-temporelle (la localisation et le temps) sont respectivement reliées à l'événement avec de nouvelles connexions rc_{when} et rc_{where} (lignes 10-11, algorithme 5). Par exemple, dans la figure 3.6, le concept *Employee* est considéré car il a une connexion sémantique rc_{who} (flèche bleue) avec l'événement. L'instance *Roland_Perrin* du concept *Employee* est sélectionnée car cet employé a quitté le bureau 413 une heure avant le déclenchement de l'événement.

Algorithme 5 Connexions ri_{what} , ri_{who} , ri_{when} et ri_{where} des instances

entrées : $O \leftarrow$ Ontologies de réseaux de capteurs et du corpus documentaire

$seuil \leftarrow$ seuil de liaison spatio-temporelle

$HeureÉvénement, LocalisationÉvénement \leftarrow$ Heure et lieu de déclenchement de l'événement

$Résultat_étape1 \leftarrow$ Les triples étiquetés rc_{what} , rc_{who} , rc_{when} et rc_{where} fournis par l'algorithme 3

sorties : Un ensemble de triplets <événement, prédicat, instance> avec ri_{what} , ri_{who} , ri_{when} et ri_{where} comme prédicats.

```

1 Résultat  $\leftarrow$  {}
2 Concepts $C_{who} \leftarrow$  ObtenirLesConcepts $C_{who}(Résultat\_étape1)$ 
3 pour chaque  $c_{who}$  de Concepts $C_{who}$  faire
4   Instances $C_{who} \leftarrow$  ObtenirLesInstances( $c_{who}, O$ )
5   pour chaque  $i_{who}$  de Instances $C_{who}$  faire
6     SuiviSpatioTemporel  $\leftarrow$  ObtenirSuiviSpatioTemporel( $i_{who}$ )
7     sélectionnerPositionSpatioTemporel  $\leftarrow$  SélectionnerPositionSpatioTemporelle(SuiviSpatioTemporels, HeureÉvénement, LocalisationÉvénement, seuil)
8     si sélectionnerPositionSpatioTemporel NOT Null alors
9       Résultat  $\leftarrow$  Résultat  $\cup$  <événement.id,  $ri_{who}, i_{who}$ >
10      Résultat  $\leftarrow$  Résultat  $\cup$  <événement.id,  $ri_{when}, sélectionnerPositionSpatioTemporel.time$ >
11      Résultat  $\leftarrow$  Résultat  $\cup$  <événement.id,  $ri_{where}, sélectionnerPositionSpatioTemporel.location$ >
12    fin
13  fin
14 fin
15 Concepts $C_{what} \leftarrow$  ObtenirLesConcepts $C_{what}(Résultat\_étape1)$ 
16 pour chaque  $c_{what}$  de Concepts $C_{what}$  faire
17   Instances $C_{what} \leftarrow$  ObtenirLesInstances( $c_{what}, O$ )
18   pour chaque  $i_{what}$  de WhatInstances faire
19     SuiviSpatioTemporel $_{what} \leftarrow$  ObtenirSuiviSpatioTemporel( $i_{what}$ )
20     sélectionnerPositionSpatioTemporel $_{what} \leftarrow$  SélectionnerPositionSpatioTemporelle(SuiviSpatioTemporels, HeureÉvénement, LocalisationÉvénement, seuil)
21     si sélectionnerPositionSpatioTemporel $_{what}$  NOT Null alors
22       Résultat  $\leftarrow$  Résultat  $\cup$  <événement.id,  $ri_{what}, i_{what}$ >
23     fin
24   fin
25 fin
26 retourner Résultat

```

La position spatio-temporelle de l'instance *Roland_Perrin* a pu être rapprochée de celle du déclenchement de l'événement grâce aux relations de liaison de données (flèches vertes) entre les instances *Office413* de l'ontologie réseau de capteurs et *Office413* de l'ontologie ressources humaines, et entre les instances *28/08/13 20:00:00* de l'ontologie réseau de capteurs et *28/08/13 18:32:50* de l'ontologie ressources humaines. Par conséquent, une nouvelle connexion étiquetée ri_{who} au niveau des instances (flèche orange) est construite entre l'événement *light wastage* et l'instance filtrée *Roland_Perrin*. Bien entendu, les connexions ri_{when} et ri_{where} sont également établies entre l'événement *light wastage*, *Office 413* et l'instance *28/08/13 18:32:50* respectivement. Dans la figure 3.6, nous ne montrons pas

les connexions ri_{when} et ri_{where} pour des raisons de visibilité. De toute évidence, dans la figure 3.6, nous ne créons pas d'autres connexions ri_{who} étant donné que les autres employés n'étaient pas dans le bon bureau au bon moment.

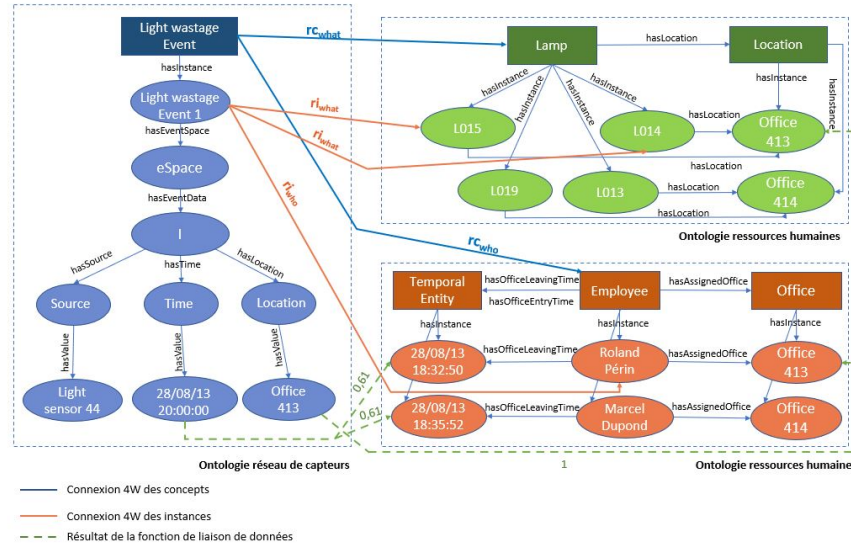


FIGURE 3.6 – Un exemple des connexions 4W des instances

— **Construction d'interconnexions étiquetées ri_{what}**

Pour construire les connexions sémantiques ri_{what} au niveau des instances, l'approche est similaire aux connexions sémantiques ri_{who} présentées précédemment. Les instances des concepts précédemment interconnectés à l'événement, par le biais des connexions sémantiques rc_{what} sont analysées itérativement (lignes 15-17, algorithme 5). Ensuite, celles dont le lieu et l'heure sont proches du déclenchement de l'événement sont sélectionnées (lignes 18-21, algorithme 5). Pour ce faire, les techniques de liaison de données sont utilisées pour comparer l'historique des positions des instances des concepts rc_{what} au lieu et à l'heure du déclenchement de l'événement (ligne 20, algorithme 5). Si la fonction de liaison de données identifie une position d'une instance comme similaire à celle du déclenchement de l'événement (ligne 21, algorithme 5), l'instance est connectée à l'événement avec une nouvelle relation étiquetée ri_{what} (ligne 22, algorithme 5).

Par exemple, dans la figure 3.6, le concept *Lamp* est considéré car il a une connexion sémantique rc_{what} (flèche bleue) avec l'événement. Les deux instances *L014* et *L015* du concept *Lamp* sont sélectionnées car elles sont situées dans *Office 413* là où l'événement s'est déclenché (flèche verte). Par conséquent, une nouvelle connexion étiquetée ri_{what} au niveau des instances (flèche orange) est construite entre l'événement *light wastage* et les instances filtrées *L014* et *L015*. Bien entendu, les instances *L013* et *L019*

du concept *Lamp* ne sont pas considérées puisqu'elles ne sont pas situées au même endroit que l'événement. Nous rappelons ici que pour établir les connexions sémantiques ri_{who} et ri_{what} , nous nous sommes basés sur l'information spatio-temporelle de l'événement déclenché. Cette l'information provient des données de déclenchement de l'événement (attribut *I*, déf. 13).

Après avoir détaillé notre démarche de construction d'interconnexions étiquetées ri_{what} , ri_{who} et ri_{when} et ri_{where} , nous expliquons dans ce qui suit comment à partir de ces interconnexions nous allons construire l'explication complète de l'événement (déf. 16).

Algorithme 6 Construction des tuples Why_i de l'explication complète

entrées : $O \leftarrow$ Ontologies réseau de capteurs et du corpus documentaire

$TuplesWhy_c \leftarrow$ Les tuples Why_c fournis par l'algorithme 4

$Résultat_étape2 \leftarrow$ Les triples étiquetés ri_{what} , ri_{who} , ri_{when} et ri_{where} fournis par l'algorithme

5

sorties : Un ensemble de tuples Why_i de l'explication complète

```

1  Résultat  $\leftarrow$  {}
2  pour chaque  $t_{Why_c}$  de  $TuplesWhy_c$  faire
3       $ConceptsC_{who} \leftarrow t_{Why_c}.C_{who}$ 
4       $InstancesC_{who} \leftarrow$  ObtenirInstances( $ConceptsC_{who}$ ,  $Résultat\_étape2$ )
5      /* faire la même chose pour obtenir les instances  $C_{what}$ ,  $C_{when}$  et  $C_{where}$  */
6      pour chaque  $instanceC_{who}$  de  $InstancesC_{who}$  faire
7           $t_{Why_i} \leftarrow TupleWhy_i(t_{Why_c})$ 
8           $t_{Why_i}.i_{who} \leftarrow instanceC_{who}$ 
9           $t_{Why_i}.i_{what} \leftarrow \{\}$ 
10          $t_{Why_i}.i_{when} \leftarrow \{\}$ 
11          $t_{Why_i}.i_{where} \leftarrow \{\}$ 
12          $VoisinsInstanceC_{who} \leftarrow$  ObtenirLesInstancesVoisines( $instanceC_{who}$ )
13         pour chaque  $instanceC_{what}$  de  $InstancesC_{what}$  faire
14             si  $instanceC_{what} \in VoisinsInstanceC_{who}$  alors
15                  $t_{Why_i}.i_{what} \leftarrow t_{Why_i}.i_{what} \cup instanceC_{what}$ 
16                  $PropriétéReliantC_{what}C_{who} \leftarrow$  ObtenirPropriétéReliant( $instanceC_{who}$ ,  $instanceC_{what}$ )
17                  $t_{Why_i}.i_{how} \leftarrow t_{Why_i}.i_{how} \cup \langle instanceC_{who}, PropriétéReliantC_{what}C_{who}, instanceC_{what} \rangle$ 
18             fin
19         fin
20         pour chaque  $instanceC_{when}$  de  $InstancesC_{when}$  faire
21             si  $instanceC_{when} \in VoisinsConceptC_{who}$  alors
22                  $t_{Why_i}.i_{when} \leftarrow t_{Why_i}.i_{when} \cup instanceC_{when}$ 
23             fin
24         fin
25         /* faire la même chose pour  $InstancesWhere$  */
26          $Résultat \leftarrow Résultat \cup t_{Why_i}$ 
27     fin
28 retourner Résultat
    
```

>>> Construction d'explication complète de l'événement

Comme nous l'avons indiqué précédemment (section 3.4.2), l'explication complète (déf. 16) de l'événement est constituée de l'explication standard (déf. 15) en plus d'un attribut Why_i . Le champ Why_i détaille les informations sur les instances reliées au déclenchement de l'événement. Il est constitué de plusieurs tuples Why_i . Dans cette section, nous expliquons comment à partir des interconnexions sémantiques ri_{what} , ri_{who} , ri_{when} et ri_{where} , nous allons pouvoir renseigner les tuples Why_i , ainsi construire l'explication complète e_c de l'événement (e_c , figure 3.2). Ce processus est détaillé dans l'algorithme 6.

L'algorithme 6 parcourt itérativement les tuples Why_c fournis par l'étape (2) du processus ISEE (algorithme 4). Pour chaque tuple Why_c , les instances du concept contenu dans l'attribut c_{who} du tuple sont obtenues à partir du résultat de l'algorithme 5 (ligne 2-4 de l'algorithme 6). Par exemple, il pourrait s'agir de l'instance *Roland_Perrin* du concept *Employee* (tuple Why_{c1} , figure 3.5). Pour chacune de ces instances l'algorithme construit un tuple Why_i avec celle-ci comme attribut i_{who} (ligne 5-7, algorithme 6). Ensuite, pour renseigner les attributs i_{what} , i_{how} , i_{when} et i_{where} du tuple Why_i , l'algorithme s'appuie sur les concepts contenus dans les champs c_{what} , c_{how} , c_{when} et c_{where} du tuple Why_c . L'algorithme parcourt itérativement les instances des concepts c_{what} du tuple Why_c (ligne 12 de l'algorithme 6), si une des instances des concepts c_{what} est reliée à l'instance i_{who} , celle-ci est ajoutée à l'attribut i_{what} du tuple Why_i (ligne 11-14, algorithme 6). La même procédure est établie pour les attributs i_{when} et i_{where} (ligne 19-23, algorithme 6). Enfin, pour renseigner l'attribut i_{how} du tuple Why_i , les relations sémantiques reliant les instances i_{what} et i_{who} sont utilisées pour construire les triplets. Nous rappelons ici que l'attribut i_{how} a pour rôle d'expliquer à l'utilisateur la relation entre l'entité responsable du déclenchement de l'événement (attribut i_{who}) et l'élément qui a permis de l'identifier comme responsable (attribut i_{how}). Par exemple, les triplets $\langle Roland_Perrin, turnsOn, L015 \rangle$ et $\langle Roland_Perrin, turnsOff, L015 \rangle$ expliquent la relation entre l'attribut i_{who} *Roland_Perrin* et l'attribut i_{what} *L015* (l'employé *Roland_Perrin* est celui qui a laissé la lampe *L015* allumée). La figure 3.7 montre l'explication complète construite à partir de l'exemple cité précédemment ainsi qu'un autre tuple Why_i qui concerne un projecteur installé dans l'environnement. Comme vous pouvez le constater le champ Why_i contient deux tuples Why_i . Certains attributs du tuple Why_{i2} ne sont pas renseignés étant donné que nous n'avons pas pu trouver des informations dans les ontologies qui nous permettent de les renseigner. Enfin, la métrique de calcul de score sera détaillée dans la section 3.5.3.3.

Après avoir détaillé les étapes (1) et (2) du processus ISEE (figure 3.1), nous passons maintenant à la dernière étape qui consiste à classer les tuples Why_c (explication standard, figure 3.5) et Why_i (explication complète, figure 3.7).

What	Luminosity						
Who	LightSensor44						
When	28/08/13 20:00:00						
Where	Office 413						
How	< 36145, lightSensor44, 28/08/13 19:00:00, Office_413, 62 >, ..., < 36191, lightSensor44, 28/08/13 20:00:00, Office_413, 68 >						
Why _i	<table border="0" style="width: 100%;"> <tr> <td style="border: 1px solid black; padding: 5px; width: 50%;"> <pre> i_What : Lamp/L014 , Lamp/L015 i_Who : Employee/ Roland Perrin i_When : TemporalEntity/28/08/13 18:32:50 i_Where : Office/Office413 i_How : < John Smith, turnsOn, L014 > < John Smith, turnsOff, L014 > < John Smith, turnsOn, L015 > < John Smith, turnsOff, L015 > Score_{RolandPerrin} = 0,81 </pre> <p style="text-align: center;">Tuple Why_{i1}</p> </td> <td style="border: 1px solid black; padding: 5px; width: 50%;"> <pre> i_What : Lamp/L48 , Brightness/ 3000 (lumens) i_Who : Projector/Projector15 i_When : - i_Where : Office/Office413 i_How : < Projector15, hasBrightness, 3000 > < Projector15, equippedWith, L48 > Score_{Projector15} = 0,61 </pre> <p style="text-align: center;">Tuple Why_{i2}</p> </td> <td style="vertical-align: middle; text-align: center;">...</td> </tr> <tr> <td colspan="3" style="text-align: center;">Tuples Why_i</td> </tr> </table>	<pre> i_What : Lamp/L014 , Lamp/L015 i_Who : Employee/ Roland Perrin i_When : TemporalEntity/28/08/13 18:32:50 i_Where : Office/Office413 i_How : < John Smith, turnsOn, L014 > < John Smith, turnsOff, L014 > < John Smith, turnsOn, L015 > < John Smith, turnsOff, L015 > Score_{RolandPerrin} = 0,81 </pre> <p style="text-align: center;">Tuple Why_{i1}</p>	<pre> i_What : Lamp/L48 , Brightness/ 3000 (lumens) i_Who : Projector/Projector15 i_When : - i_Where : Office/Office413 i_How : < Projector15, hasBrightness, 3000 > < Projector15, equippedWith, L48 > Score_{Projector15} = 0,61 </pre> <p style="text-align: center;">Tuple Why_{i2}</p>	...	Tuples Why _i		
<pre> i_What : Lamp/L014 , Lamp/L015 i_Who : Employee/ Roland Perrin i_When : TemporalEntity/28/08/13 18:32:50 i_Where : Office/Office413 i_How : < John Smith, turnsOn, L014 > < John Smith, turnsOff, L014 > < John Smith, turnsOn, L015 > < John Smith, turnsOff, L015 > Score_{RolandPerrin} = 0,81 </pre> <p style="text-align: center;">Tuple Why_{i1}</p>	<pre> i_What : Lamp/L48 , Brightness/ 3000 (lumens) i_Who : Projector/Projector15 i_When : - i_Where : Office/Office413 i_How : < Projector15, hasBrightness, 3000 > < Projector15, equippedWith, L48 > Score_{Projector15} = 0,61 </pre> <p style="text-align: center;">Tuple Why_{i2}</p>	...					
Tuples Why _i							

FIGURE 3.7 – Un exemple d’explication complète de l’événement gaspillage de lumière

3.5.3.3 Un processus pour le classement des explications

Dans cette section, nous présentons **la contribution 2.2** (section 1.3.2), un processus pour le classement des explications. Ce dernier constitue l’étape (3) du processus ISEE (figure 3.1). Il a pour but de calculer les scores des tuples Why_c de l’explication standard (e_s , figure 3.2) et Why_i de l’explication complète (e_c , figure 3.2).

Nous présentons le processus de calcul de score des tuples Why_c de l’explication standard (e_s , figure 3.2) dans l’algorithme 7. Ce dernier parcourt itérativement chaque tuple Why_c puis calcule son score en se basant sur une fonction de calcul de score (ligne 3, algorithme 7). Le même processus est appliqué pour les tuples Why_i de l’explication complète (e_c , figure 3.2) sauf que nous employons différentes fonctions de calcul de score. Nous détaillons ces fonctions dans ce qui suit.

Algorithme 7 Classement des tuples Why_c

entrées : *RésultatAlgorithme4* ← *Tuples Why_c produites par l’algorithme 4*

événement ← *l’événement déclenché*

- 1 *TuplesWhy_c* ← *ObtenirTuplesWhy_c(RésultatAlgorithme4)*
 - 2 **pour chaque** t_{why_c} **de** *TuplesWhy_c* **faire**
 - 3 $t_{why_c}.$ Score ← *ObtenirScore(event, t_{why_c})*
 - 4 **fin**
-

- **Calcul de score des tuples Why_c de l’explication standard** : la fonction de calcul de score exploite les informations dont nous disposons sur les tuples Why_c . Elle est basée sur deux critères d’évaluation, à savoir **la complétude** (*comp*) et le

score de confiance (*conf*). La complétude désigne le nombre d'éléments 4W1H renseignés par le tuple Why_c . Ce critère découle de l'idée que plus les éléments de réponse 4W1H sont renseignés par le tuple Why_c , plus celui-ci est crédible. Le **score de confiance** (*conf*) représente la moyenne des scores de confiance des alignements utilisés pour construire le tuple Why_c . Ces scores sont calculés par l'algorithme d'alignement dans la deuxième étape du processus d'intégration des connaissances de domaine (section 3.5.1.2). Ces alignements constituent le point de départ du processus d'explication d'événements et toutes les interconnexions établies ensuite sont basées dessus. Nous en avons donc conclu que plus le score de confiance d'une relation d'alignement est élevé, plus les interconnexions établies sur la base de celle-ci sont fiables. Nous proposons donc la formule suivante :

$$\text{Score}_{\text{tuple}Why_c} = \frac{1}{\alpha + \beta} * (\alpha \cdot \text{comp} + \beta \cdot \text{conf})$$

- α, β sont des coefficients de pondération, elles équilibrent l'importance accordée à chacun des deux critères d'évaluation.
- *comp* dénote la complétude du tuple Why_c , $\text{comp} = \frac{m}{5}$, m est le nombre de dimensions 4W1H renseignées tuple Why_c .
- *conf* est la moyenne des scores de confiance des alignements qui ont été utilisés pour construire le tuple Why_c (section 3.5.1.2), $\text{conf} = \frac{1}{p} \sum_{i=1}^p c_i$, c_i est le score de confiance du i ème alignement.

Par exemple, si nous reprenons l'explication standard présentée dans la figure 3.5 et que nous fixons les valeurs de α et β à 0.5, les scores des deux tuples (tuple Why_{c1} et tuple Why_{c2}) sont calculés comme suit :

$$\text{Score}_{\text{Tuple}Why_{c1}} = 0.5 * (5/5) + 0.5 * 0.76 = 0.88$$

$$\text{Score}_{\text{Tuple}Why_{c2}} = 0.5 * (3/5) + 0.5 * 0.71 = 0.65$$

Dans le tuple Why_{c1} , la complétude est égale à 5/5 vu que tous les attributs du tuple sont renseignés. Le score de confiance est égal à 0.76, la moyenne des scores de confiance des alignements qui ont servi à construire le tuple Why_{c1} (cf. figure 3.3). Dans le tuple Why_{c2} , la complétude est égale à 3/5 vu que seulement trois des attributs du tuple sont renseignés. La moyenne des scores de confiance est égale à 0.71 (cf. figure 3.3).

- **Calcul de score des tuples Why_i de l'explication complète** : la fonction de calcul de score des tuples Why_i s'appuie sur deux critères d'évaluation supplémentaires, à savoir la proximité spatio-temporelle (*prox*) et la diversité (*diver*). Nous avons choisi de nous appuyer sur le critère de la proximité spatio-temporelle car nous considérons que c'est un aspect très important dans notre contexte. **La proximité spatio-temporelle** (*prox*) désigne la distance en terme de temps et d'espace entre deux objets [79]. Nous faisons l'hypothèse que les entités de l'environnement qui

ont une position proche du lieu et de l'heure du déclenchement de l'événement ont plus de chance d'être liées à son déclenchement comparé à d'autres entités. La diversité compte le nombre d'instances *What* connectées à l'événement. En effet, vu que les instances *What* expliquent la relation entre l'événement et l'entité responsable de son déclenchement), nous avons estimé que plus il y a d'entités qui relient le tuple à l'événement, plus la relation de ce tuple est forte avec le dernier. Nous proposons donc la formule suivante :

$$\text{Score}_{\text{tuple}Why_i} = \frac{1}{\alpha + \beta + \gamma + \delta} * (\alpha \cdot \text{comp} + \beta \cdot \text{conf} + \gamma \cdot \text{diver} + \delta \cdot \text{prox})$$

- $\alpha, \beta, \gamma, \delta$ sont des coefficients de pondération.
- *comp* dénote la complétude du tuple *Why_i*, $\text{comp} = \frac{m}{5}$, m est le nombre de dimensions 4W1H renseignées par tuple *Why_i*.
- *conf* est la moyenne des scores de confiance des alignements qui ont été utilisés pour construire le tuple *Why_i* (section 3.5.1.2), $\text{conf} = \frac{1}{p} \sum_{i=1}^p c_i$, c_i est le score de confiance du ième alignement.
- *diver* fait référence à la diversité des instances *What*, $\text{diver} = \frac{n}{I}$, n est le nombre d'instances *What* reliant le tuple *Why_i* à l'événement et I est le nombre total d'instances *What*.
- *prox* désigne la proximité spatio-temporelle du tuple *Why_i*

$$\text{prox}(i_{where}, i_{when}) = \begin{cases} 0 & \text{if } i_{when} > \text{event}.t \\ 0.25 & \text{if } i_{when} \text{ and } i_{where} \text{ are empty} \\ 0.75 & \text{if } i_{when} < \text{event}.t \text{ and } i_{where} \text{ is empty} \\ 1 & \text{if } i_{where} \leq \text{event}.l \text{ and } i_{when} < \text{event}.t \end{cases}$$

i_{where} et i_{when} sont les instances *When* et *Where* du tuple *Why_i*, *event.t* et *event.l* sont respectivement l'heure et le lieu de déclenchement de l'événement.

Nous considérons ici que pour qu'une entité soit liée au déclenchement d'un événement, elle doit avoir une position spatiale proche du lieu du déclenchement de l'événement (d'où la condition $i_{where} \leq \text{event}.l$ dans *prox*) et une position temporelle qui précède l'heure du déclenchement de l'événement (si elle est renseignée), étant donné qu'une entité ne peut pas être liée au déclenchement d'un événement si celle-ci entre à l'environnement après son déclenchement (d'où la condition $i_{when} < \text{event}.t$ dans *prox*).

Par exemple, si nous reprenons l'explication complète présentée dans la figure 3.7 et que nous fixons les valeurs des coefficients de pondération à 0.25, le score du tuple *Why_{i1}* est calculé comme suit :

$$\text{Score}_{\text{Tuple}Why_{i1}} = 0.25 * (5/5) + 0.25 * 0.76 + 0.25 * (2/4) + 0.25 * 1 = 0.815$$

La complétude est égale à 5/5 vu que tous les attributs du tuple sont renseignés. Le score de confiance est égale à 0.76, la moyenne des scores de confiance des alignements

qui ont servi à construire le tuple Why_{i1} (cf. figure 3.3). La diversité est égale à $2/4$ si nous nous tenons juste au deux tuples Why_i présentés dans la figure 3.7 (il y a 4 instances *What* dont deux appartiennent au tuple Why_{i1}). Enfin, la proximité spatio-temporelle est égale à 1 puisque l'attribut *Where* est similaire au lieu du déclenchement de l'événement (*Office413*), tandis que l'attribut *When* renseigne un horaire qui précède l'heure du déclenchement de l'événement. Notez ici que dans les expérimentations menées pour évaluer le système ISEE (chapitre 4), nous avons fixé les coefficients de pondération à $1/4$ respectivement. L'automatisation de la définition de ces coefficients pour améliorer les performances du système sera abordée dans des travaux futurs.

3.6 Conclusion

Dans ce chapitre, nous avons présenté notre proposition, le système ISEE pour l'explication d'événements déclenchés dans les environnements hybrides. ISEE est basé à la fois sur (i) les données de réseaux de capteurs et de corpus documentaire et sur (ii) des ontologies pour modéliser sémantiquement les connaissances du domaine. Pour modéliser les événements et les intégrer dans le système ISEE, nous avons proposé un modèle multidimensionnel pour la définition des événements dans les environnements hybrides qui est utilisé ensuite pour étendre l'ontologie HSSN [116] avec la notion d'événement. Ce modèle constitue **la contribution 1.1** de cette thèse. Nous avons également proposé un modèle inspiré de l'approche 5W1H [72] pour la structuration des explications d'événements. Ce modèle décrit les résultats obtenus à chaque étape du processus ISEE, il a pour but de présenter à l'utilisateur des explications simples et faciles à comprendre. Il constitue **la contribution 1.2** (section 1.3.1).

Le système ISEE consiste en deux processus principaux, à savoir, le processus d'intégration des connaissances du domaine et le processus ISEE proprement dit. Le processus d'intégration des connaissances du domaine a pour objectif d'intégrer les informations issues de l'environnement (données de capteurs et du corpus documentaire) dans les ontologies du domaine et ensuite d'établir des alignements classiques entre ces ontologies en s'appuyant sur des outils existants. Le processus ISEE s'appuie sur le graphe de connaissances (l'ensemble des ontologies de domaine après leur instanciation et alignement) construit dans cette première étape et sur les données de définition et de déclenchement de l'événement pour établir une deuxième couche d'interconnexions sensibles au contexte entre les ontologies de domaine. Le graphe de connaissances est ensuite analysé pour construire les candidats d'explication d'événement. Le processus ISEE est composé de trois étapes. La première étape s'appuie sur les alignements classiques établis précédemment et sur les données de définition de l'évènement pour construire des interconnexions sensibles au contexte (liens *What* et *Who*) entre l'évènement et les concepts connexes des ontologies de domaine. La deuxième étape se base sur les données de déclenchement de l'évènement et établit des interconnexions sensibles au contexte (liens *What*, *Who*, *When*, et *Where*) entre l'évènement et les instances connexes des ontologies de domaine. Ces deux étapes constituent **la contribution 2.1**

de cette thèse (section 1.3.2). Elles permettent de filtrer successivement les concepts et les instances des ontologies, à des fins d'explication du déclenchement d'un événement. Selon le niveau d'urgence de la demande d'explication, les calculs seront étendus jusqu'aux instances ou limités aux concepts. Le premier filtrage réduit les ontologies aux seuls concepts sémantiquement liés à l'événement (niveau d'urgence moyen), tandis que le second filtrage réduit les ontologies à quelques instances de ces concepts sémantiquement liés à l'événement (niveau d'urgence faible). Enfin, dans la dernière étape, sur la base de toutes les interconnexions sémantiques ainsi établies, les candidats d'explication sont construits et classés par ordres de pertinence en s'appuyant sur une métrique que nous avons proposée. Cette étape constitue **la contribution 2.2** (section 1.3.2).

À notre connaissance, le système ISEE est le premier système d'explication d'événements dans les environnements hybrides composés de réseau de capteurs et de corpus documentaires hétérogènes. Afin de tester le système ISEE dans un contexte réel, nous avons développé un premier prototype que nous avons appelé ISEEapp. Ce prototype ainsi que les évaluations qui ont été conduites seront détaillées dans le chapitre suivant. Enfin, notez qu'à ce stade, nous n'exploitons pas encore le retour des utilisateurs sur les explications renvoyées par le système. Ce sujet sera traité dans des travaux futurs.

Chapitre 4

Expérimentation et Évaluation

4.1 Introduction

Dans le chapitre 3 nous avons détaillé notre proposition, ISEE un nouveau système pour l'explication des événements dans les environnements connectés. Dans ce chapitre nous présentons tout d'abord la plateforme ISEEapp qui concrétise cette proposition sous forme d'une application Web. Ensuite, nous nous consacrons à l'évaluation et à la validation de notre proposition. Pour rappel, le système ISEE est basé sur deux contributions : (i) la première contribution concerne la modélisation des événements dans les environnements hybrides. Plus précisément, nous proposons un modèle pour la définition des événements et un modèle inspiré de l'approche 5W1H pour la structuration des explications des événements; (ii) la deuxième contribution porte sur l'interconnexion ciblée des ontologies. Nous proposons un processus pour l'explication des événements en se basant sur l'interconnexion et le filtrage du graphe sémantique (l'ensemble des ontologies de domaine). Ainsi, nous conduisons dans ce chapitre trois expérimentations pour évaluer ces contributions :

1. **Évaluation du modèle de définition des événements** (contribution 1.1., section 1.3.1) : le modèle que nous proposons a pour but de permettre la définition des événements selon différents axes de description (capteurs et documents) favorisant ainsi le rapprochement entre les données du réseau de capteurs et celles du corpus de documents. Par conséquent, pour évaluer cette première contribution, nous mesurons l'évolution de la connectivité du graphe sémantique selon la diversité des événements définis avec notre modèle. Ensuite, nous comparons les résultats avec ceux obtenus en utilisant d'autres modèles de définition d'événements (section 4.3).
2. **Évaluation des résultats retournés par le processus d'explication des événements** (contribution 2., section 1.3.2) : le processus ISEE a pour objectif d'analyser les données du graphe sémantique et de construire des explications aux événements déclenchés dans l'environnement. Nous nous focalisons dans cette évaluation sur la qualité de ces explications. Nous avons tout d'abord conduit

une preuve de concept (section 4.4.1) sur un petit jeu de données simulé, en utilisant des métriques classiques de la littérature (précision, rappel, F-score) pour évaluer les résultats retournés. Ensuite, après le développement de la plateforme ISEEapp, nous avons conduit une expérimentation sur un jeu de données plus large, avec des données réelles de capteurs en s'appuyant sur des métriques que nous proposons (complétude et cohérence, section 4.4.2).

3. **Évaluation de l'interface utilisateur pour l'explication des événements** (contributions (i) et (ii)) : l'objectif de notre modèle 5W1H pour la structuration des explications (contribution 1.2., section 1.3.1) est de retourner des réponses claires et faciles à comprendre par les utilisateurs de l'environnement connecté. Notre modèle s'appuie bien évidemment sur le processus d'explication des événements (contribution 2., section 1.3.2). Cette dernière expérimentation évalue la facilité de compréhension des résultats retournés par la plateforme ISEEapp. Nous conduisons une enquête auprès d'un groupe d'utilisateurs potentiels. Ensuite, nous calculons les moyennes des niveaux de compréhension et nous discutons les résultats (section 4.5).

Le reste de ce chapitre est organisé comme suit. La section 4.2 présente l'architecture générale de la plateforme ISEEapp. Ensuite, les trois sections 4.3, 4.4 et 4.5 détaillent respectivement les trois expérimentations (1), (2) et (3). Nous concluons ce chapitre dans la section 4.6.

4.2 La plateforme ISEEapp

La plateforme ISEEapp met en œuvre toutes nos propositions présentées dans le chapitre 3. Elle est développée en langage Python en utilisant le micro cadre (framework) open-source Flask¹. La raison de ce choix réside dans le fait que le cadre Flask est très léger, simple à utiliser et intègre tous les éléments nécessaires au développement des applications Web (serveur intégré, débogueur rapide fourni, etc.). La figure 4.1 présente l'architecture générale de la plateforme ISEEapp.

L'utilisateur communique avec la plateforme à travers une interface graphique développée en HTML, CSS et JavaScript. Le dépôt de données est composé de cinq jeux de données brutes (données du réseau de capteurs, corpus documentaire, définitions des composantes de l'environnement, explications des événements passés et ontologies de domaine). Les données du réseau de capteurs, le corpus documentaire, les définitions des composantes de l'environnement et les explications des événements passés sont stockés sous forme de fichiers JSON². Nous avons choisi ce format étant donné qu'il est très communément utilisé et facile à manipuler grâce à plusieurs bibliothèques Python. L'accès aux ontologies de domaine se fait à travers leurs URI (Uniform Resource Identifier). Le backend de la plateforme ISEEapp (figure 4.1) est constitué de deux

1. <https://flask.palletsprojects.com/en/2.1.x/>

2. <https://www.json.org/json-fr.html>

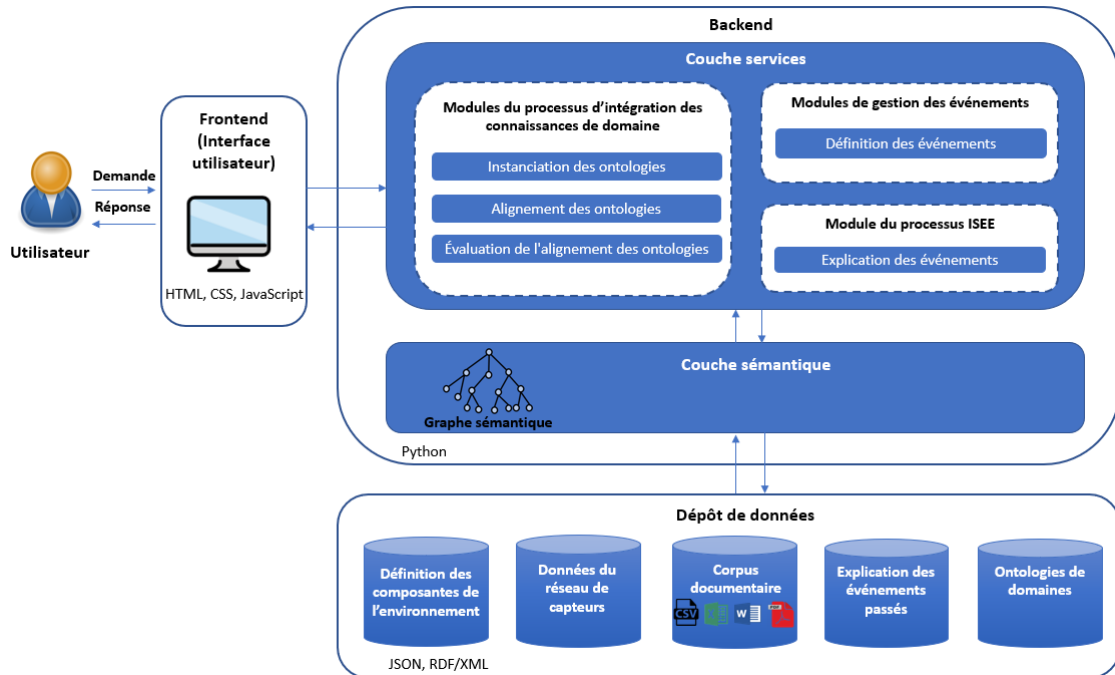


FIGURE 4.1 – Architecture de la plateforme ISEEapp

couches, à savoir **la couche services** qui comprend les modules principaux de la plateforme ISEEapp et **la couche sémantique** qui permet l'accès aux ressources du dépôt de données ainsi que la gestion du graphe sémantique. Le graphe sémantique contient l'ensemble des ontologies de domaine après leurs instanciations et leurs alignements grâce aux données de capteurs et du corpus documentaire. Il est alimenté et manipulé par les différents modules de la couche services en s'appuyant sur les ressources du dépôt de données. Il permet ainsi de faire la liaison entre les modules de la couche services et les ressources du dépôt de données. Le graphe sémantique est sérialisé en RDF/XML. Nous avons choisi RDF/XML car c'est un modèle de données largement utilisé et fiable pour représenter les données sémantiques [1].

La couche services comprend cinq modules organisés en trois catégories :

- **Intégration des connaissances de domaine** : cette catégorie correspond aux différentes étapes du processus d'intégration des connaissances de domaine (intégration des connaissances de domaine, figure 5.2). Elle comprend trois modules :
 - **Instanciation des ontologies** : l'objectif principal de ce module est d'instancier les ontologies de domaine (*ontologies de domaine*, figure 4.1) à partir du dépôt de données, le résultat de cette étape est ainsi intégré dans le graphe sémantique (*instanciation des ontologies*, figure 4.1). Pour l'instant, le processus d'instanciation n'est automatisé que pour les données de capteurs qui sont récupérées sous forme de fichier CSV et utilisées pour instancier l'ontologie HSSN-étendue (sections 2.2.2 et 2.5). L'instanciation des ontologies de domaine à partir du corpus documentaire se fait pour l'instant

manuellement. Nous analysons les documents et nous les traduisons sous forme de tableaux de données qui sont ensuite utilisés pour instancier les ontologies de domaine. L'automatisation de ce processus est traitée en détail dans la section perspectives de cette thèse (section 5.2).

- **Alignement des ontologies** : ce module a pour objectif d'aligner les ontologies de domaine en s'appuyant sur des outils existants. La version actuelle de la plateforme intègre les deux outils **AgreementMakerLight**³ et **TheAlignmentAPI**⁴. L'utilisateur a la possibilité de choisir l'outil d'alignement qui lui convient. Nous avons choisi ces outils car ils sont très communément utilisés dans la littérature, de plus, ils sont performants et faciles à mettre en œuvre. Le résultat de cette étape est intégré dans le graphe sémantique (*alignement des ontologies*, figure 4.1).
- **Évaluation de l'alignement des ontologies** : ce module intègre la formule d'évaluation des alignements d'ontologies présentée dans la section 3.5.2. Il permet à l'utilisateur d'évaluer l'alignement et d'obtenir un score qui lui indique si l'alignement établi lui permettra éventuellement d'obtenir de bonnes explications. Si ce score est trop faible, il est probable que les explications générées lors des étapes suivantes du processus ISEE soient de moindre qualité (*évaluation de l'alignement des ontologies*, figure 4.1).
- **Gestion des événements** : cette catégorie permet de gérer la modélisation des événements, elle comprend le module :
 - **définition des événements** : ce module permet à l'utilisateur de définir les événements qu'il souhaite détecter dans l'environnement à travers un formulaire. Ce dernier est basé sur le modèle de définition d'évènements présenté dans le chapitre précédent (déf. 11, 12 et 13). La nouvelle définition est ainsi ajoutée dans le graphe sémantique (*définition des événements*, figure 4.1).
- **Processus ISEE** : cette catégorie a pour but de construire et retourner les explications aux événements déclenchés dans l'environnement connecté. Elle est constituée d'un seul module :
 - **Explication des événements** : ce module englobe les trois étapes du processus ISEE (étapes (1), (2) et (3), figure 3.1). Il permet à l'utilisateur de choisir l'évènement qu'il souhaite expliquer, calcule un ensemble d'explications potentielles en s'appuyant sur les algorithmes proposés (section 3.5.3) et les retourne à l'utilisateur (*explication des événements*, figure 4.1).

Pour implémenter ces différents modules en langage python, nous nous sommes principalement basés sur : (i) le paquet *json*⁵ pour manipuler les fichiers JSON du dépôt

3. <https://github.com/AgreementMakerLight/AML-Project>

4. <https://moex.gitlabpages.inria.fr/alignapi/>

5. <https://docs.python.org/3/library/json.html>

de données; (ii) le paquet *Owlready2*⁶ pour charger les ontologies en tant qu'objets Python, les modifier et les sauvegarder; (iii) le module *re*⁷ pour analyser les fichiers retournés par l'outil d'alignement en utilisant les expressions régulières. Les lecteurs qui souhaitent avoir plus de détails sur le développement de la plateforme ISEEapp sont invités à consulter l'annexe A de ce manuscrit.

Notez ici que cette première version de l'application ISEEapp a pour but principal d'implémenter et de tester le système ISEE. Par conséquent, la création des profils utilisateurs (administrateur, employé, etc.) permettant d'accéder à différents niveaux de fonctionnalités sera abordée dans des travaux futurs.

Après avoir présenté l'architecture générale de la plateforme ISEEapp et ses différentes composantes, nous utilisons cette plateforme pour procéder aux évaluations de nos contributions. Nous commençons dans la section suivante par l'évaluation de notre modèle de définition des événements (contribution 1.1., section 1.3.1).

4.3 Évaluation du modèle de définition des événements

Dans cette section, nous conduisons une expérimentation qui a pour objectif d'évaluer notre modèle de définition des événements (contribution 1.1., section 1.3.1). Concrètement, nous souhaitons savoir si le modèle de définition des événements proposé permet réellement de rapprocher les données du réseau de capteurs et les données du corpus documentaire. Pour ce faire, nous mesurons **la connectivité** du graphe sémantique (l'ensemble des ontologies de domaine après leurs alignements) en fonction du nombre d'événements qui y sont définis. Ensuite, nous comparons les résultats avec ceux obtenus avec d'autres modèles de définition des événements. Dans ce qui suit, nous détaillons le protocole expérimental dans la section 4.3.1. Nous présentons et discutons les résultats dans la section 4.3.2.

4.3.1 Protocole expérimental

Pour évaluer notre modèle de définition des événements, Nous avons choisi le scénario d'un grand bâtiment de recherche dans un campus universitaire. Nous reprendrons ce même exemple plus tard dans ce chapitre (sections 4.4 et 4.5).

4.3.1.1 Jeu de données

Le graphe sémantique représentant le bâtiment de recherche est composé de plusieurs ontologies pour modéliser le corpus de documents, les données du réseau de capteurs et les événements. Nous ne procédons pas à l'instanciation de ces ontologies dans cette expérimentation étant donné que la définition des événements et l'alignement des ontologies de domaine se font au niveau conceptuel. Dans ce qui suit, nous

6. <https://owlready2.readthedocs.io/en/v0.35/index.html>

7. <https://docs.python.org/3/library/re.html>

détaillons le choix des ontologies pour modéliser les données du réseau de capteurs, le corpus documentaire et les événements.

— **Modélisation des données de capteurs**

Comme nous l’avons expliqué précédemment dans la section 2.5, nous choisissons l’ontologie HSSN [116] pour modéliser le réseau de capteurs. Toutefois, nous ne l’étendons pas ici avec notre modèle de définition des événements. Nous procédons ainsi parce que nous souhaitons dans cette expérimentation nous comparer à des ontologies permettant de définir des événements. Ces ontologies sont dédiées uniquement à la modélisation des événements et ne comportent pas de concepts ou de propriétés permettant de modéliser les réseaux de capteurs. Par conséquent, nous ne pouvons pas les comparer à l’ontologie HSSN-étendue. Pour résoudre ce problème, nous utilisons l’ontologie HSSN dans sa version originale pour modéliser uniquement les données du réseau de capteurs. La modélisation des événements sera détaillée un peu plus loin dans cette section.

— **Modélisation du corpus de documents**

Pour modéliser les données du corpus de documents, nous avons choisi les deux ontologies **SAREF4BLDG** [117] et **EOSK** [195] pour modéliser respectivement les données du domaine du bâtiment et des ressources humaines. La raison de ce choix réside dans le fait que ces deux ontologies couvrent l’ensemble des concepts et des relations sémantiques nécessaires à la représentation de l’environnement connecté (le bâtiment de recherche).

— **Modélisation des événements**

Pour modéliser les événements définis dans l’environnement connecté, nous implémentons notre modèle de définition des événements sous forme d’une ontologie en utilisant l’éditeur open-source *Protégé*⁸. Nous appelons cette ontologie **EDOHE** (an Event Description Ontology for Heterogeneous Environments). Afin de comparer notre modèle de définition à d’autres travaux, nous choisissons les deux ontologies **SEM** [176] (Simple Event Model) et **LODE** [158] (Linking Open Descriptions of Events). Ce choix se justifie par le fait que ces ontologies sont communément utilisées dans la littérature. Elles sont facilement accessibles en ligne. De plus, les ontologies **SEM** et **LODE** emploient un vocabulaire simple et générique (par exemple, Event, Time, Location, InvolvedObject, etc.). Elles peuvent donc être utilisées pour définir n’importe quel type d’événements.

Après avoir présenté le jeu de données utilisé pour l’évaluation de notre modèle de définition d’événements, nous détaillons dans la section suivante la métrique utilisée pour mesurer la connectivité du graphe sémantique et la mise en œuvre de l’expérimentation.

8. <https://protege.stanford.edu/>

4.3.1.2 Métriques et mise en œuvre

Dans cette section nous présentons tout d’abord la métrique que nous utilisons pour calculer la connectivité du graphe sémantique. Ensuite, nous détaillons comment nous avons procédé pour exploiter cette métrique dans le graphe sémantique.

— Métriques d’évaluation

Pour mesurer la connectivité du graphe sémantique, nous n’avons pas pu trouver, dans la littérature, une métrique adaptée à notre contexte. Étant donné que nous avons déjà proposé une métrique pour calculer le score de connectivité d’un alignement (section 3.5.2), nous ré-exploitions cette métrique dans le cadre de cette expérimentation. Nous calculons le score de connectivité du graphe sémantique qui utilise notre modèle de définition des événements et nous le comparons aux scores de connectivité obtenus dans des graphes sémantiques utilisant d’autres modèles de définition des événements. Bien entendu, dans les deux cas, nous utilisons le même outil d’alignement et les mêmes ontologies de domaine. Le fait d’obtenir un meilleur score de connectivité dans le graphe sémantique utilisant notre modèle de définition des événements nous permettra d’affirmer que ce dernier favorise le rapprochement des différentes ontologies comparé aux autres modèles de définition des événements. En guise de rappel la métrique de connectivité est calculée comme suit :

$$\text{Connectivité} = \frac{1}{\alpha + \beta + \gamma} (\alpha \cdot \text{Onto} + \beta \cdot \text{Con} + \gamma \cdot \text{Conf})$$

— *Onto* : la proportion des ontologies alignées

$$\text{Onto} = \frac{\text{Nombre d'ontologies alignées}}{\text{Nombre total d'ontologies}}$$

— *Con* : la proportion des concepts alignés.

$$\text{Con} = \frac{\text{Nombre de concepts alignés}}{\text{Nombre total de concepts}}$$

— *Conf* : dénote la moyenne des scores de confiance des correspondances

$$\text{Conf} = \frac{1}{n} \sum_{i=1}^n c_i, \quad n \text{ est le nombre total de correspondances et } c_i \text{ le score de confiance de la } i\text{ème correspondance.}$$

Nous fixons dans cette expérimentation les coefficients α , β et γ à 1/3.

— Mise en œuvre

Nous mesurons la connectivité du graphe sémantique quatre fois : sans définition d’évènements, avec 10, 30 et 40 définitions d’évènement. Dans ces quatre cas de figure, nous définissons tout d’abord les événements sur l’ontologie EDOHE (par exemple, gaspillage de lumière, haut niveau de CO₂, haut niveau d’humidité, etc.). Ensuite, nous utilisons l’outil d’alignement *AgreementMakerLight* [52] pour aligner les différentes ontologies (EDOHE, HSSN, SAREF4BLDG et EOSK). Enfin, nous calculons la connectivité du graphe sémantique. Par la suite, nous répétons la même procédure avec les deux autres ontologies de modélisation d’évènement **SEM** et **LODE**. Nous avons choisi l’outil d’alignement *AgreementMakerLight*, parce qu’il est performant, facilement configurable et simple d’utilisation. Il dispose

d'une interface utilisateur qui permet de choisir les ontologies à aligner, de modifier la configuration de l'alignement et de corriger, si on le souhaite, l'alignement produit. Pour calculer la connectivité du graphe sémantique, nous avons implémenté une fonction python qui prend en entrée : (i) les ontologies de domaine et (ii) les fichiers d'alignements retournés par l'outil *AgreementMakerLight*, ensuite, elle calcule et retourne le score de la connectivité du graphe sémantique.

Après avoir présenté le jeu de données, la métrique pour le calcul de connectivité et la mise en œuvre (contribution 1.1., section 1.3.1), dans la section suivante, nous présentons les résultats de l'évaluation (contribution 2., section 1.3.2).

4.3.2 Résultats

La figure 4.2 représente la courbe d'évolution de la connectivité du graphe sémantique selon le nombre d'événements définis dans les trois ontologies EDOHE, LOD [158] et SEM [176]. La connectivité est passée de 0.67 avant la définition des événements à 0.73 après la définition de 30 événements pour le graphe sémantique avec l'ontologie EDOHE. La connectivité des deux graphes sémantiques utilisant les ontologies SEM et LOD a aussi augmenté, de 0.66 à 0.69 pour l'ontologie SEM, et de 0.64 à 0.67 pour l'ontologie LOD. Toutefois, le taux d'amélioration de la connectivité du graphe sémantique avec l'ontologie EDOHE (de 0.67 à 0.73 donc 0.06 de taux d'amélioration) est meilleur que ceux des deux graphes sémantiques utilisant les ontologies SEM (0.03) et LOD (0.03). Ceci peut être expliqué par le fait que notre modèle de définition des événements intégré dans l'ontologie EDOHE offre la possibilité de décrire les événements en utilisant un ensemble de concepts beaucoup plus riche (quatre dimensions, Source, Feature, Time et Location, en plus de quatre éléments par dimension OriginConcept, RelatedConcept, RelatedRelation et Constraints). Nous remarquons également qu'en dépassant 30 événements, la connectivité du réseau ne s'améliore presque plus, cela peut s'expliquer par le fait qu'en dépassant un certain nombre d'événements, les nouveaux événements définis ne contribuent plus à l'amélioration de la connectivité du graphe sémantique car ils utilisent les mêmes concepts que ceux utilisés auparavant pour définir les nouveaux événements.

Pour résumer, les résultats de l'évaluation de la connectivité nous ont permis de valider notre modèle de définition des événements. La connectivité est passée de 0.67 à 0.73 après la définition de 30 événements, ceci démontre que le modèle de définition des événements que nous proposons permet de rapprocher les données des capteurs et les données du corpus documentaire à travers le graphe sémantique.

À présent, après avoir présenté l'évaluation du modèle de définition des événements, la section suivante est consacrée à l'évaluation des résultats retournés par le processus d'explication des événements.

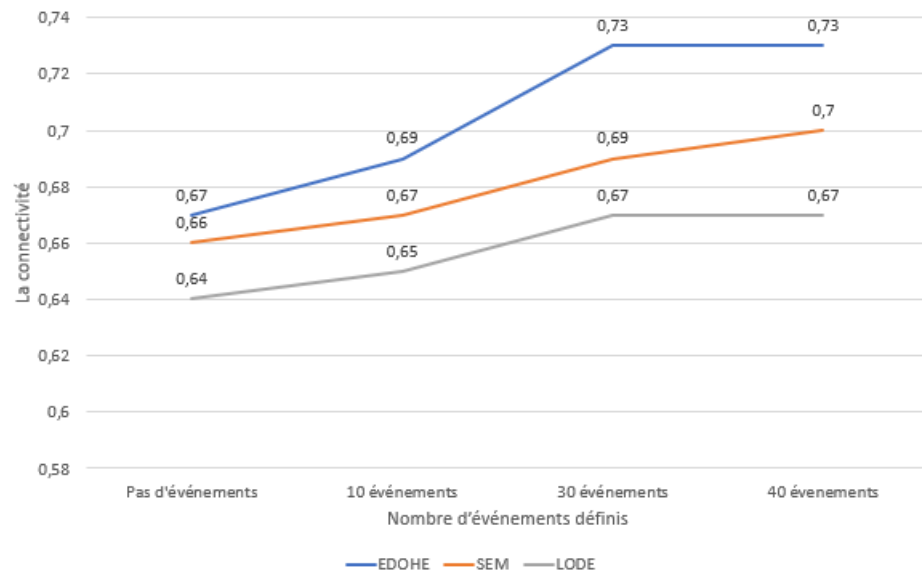


FIGURE 4.2 – Évolution de la connectivité en fonction du nombre d'événements défini pour les trois ontologies EDOHE, LODE et SEM

4.4 Évaluation des résultats retournés par le processus d'explication des événements

Dans cette section, nous évaluons les réponses retournées par le processus d'explication des événements. Nous conduisons deux expérimentations : (i) une preuve de concept manuelle sur des données simulées dans le contexte d'un parking connecté (section 4.4.1) et (ii) une expérimentation sur la plateforme ISEEapp avec des données de capteurs réelles dans le contexte du bâtiment de recherche mentionné dans la section précédente (section 4.4.2).

Notez ici que pour que les expérimentations soient conduites dans des conditions réelles, nous avons d'abord tenté de trouver un jeu de données composé de données de capteurs et du corpus documentaire correspondant. Malheureusement, nous n'avons pas pu trouver sur internet un jeu de données qui satisfait cette condition, et encore moins, un jeu de données avec un ensemble d'événements et leurs explications respectives. Dans ce qui suit, nous décrirons comment nous avons traité ce problème pour les deux expérimentations.

- **Preuve de concept** : dans cette expérimentation, les données de capteurs ainsi que l'instance de l'événement à expliquer ont été simulées. Le but étant ici de conduire une preuve de concept, l'instanciation des ontologies de domaine a été ajustée manuellement en ajoutant l'explication que le système doit retourner. Le fait de connaître cette explication, nous a permis d'évaluer le système avec les métriques classiques des systèmes question-réponse (précision, rappel, F-score)

et de valider dans un premier temps la preuve de concept.

- **Expérimentation sur la plateforme ISEEapp** : dans cette expérimentation, les données du réseau de capteurs sont des données réelles. Les événements ont été également détectés dans le jeu de données et ne sont pas simulés. Toutefois, à moins que nous soyons physiquement dans l'environnement d'où proviennent les données, nous n'avons aucun moyen de savoir les vrais raisons du déclenchement de ces événements. Dans ce cas, nous avons le choix entre deux approches : (i) ajuster les instances des ontologies de domaine en mettant pour chaque événement détecté une ou plusieurs explications; (ii) mener un processus d'instanciation automatique des ontologies de domaine (par exemple, attribuer à chaque salle, un système de ventilation et d'éclairage, attribuer aux employés, un temps d'entrée et de sortie respectivement entre 07h et 10h et entre 16h et 19h) en s'assurant que le résultat soit le plus proche possible de la réalité, sans apporter aucun ajustement ou modification en relation avec les événements déclenchés.

La première option (i) a deux inconvénients : tout d'abord, vu que la taille du jeu de données est conséquente (Section 4.4.2.1), l'ajustement des ontologies de domaine pour établir manuellement plusieurs candidats d'explication pour chaque événement détecté est très fastidieux et nécessite beaucoup de temps. De plus, le fait d'ajuster l'ensemble des ontologies implique que nous n'allons pas évaluer le système dans un contexte réel où l'on est supposé n'avoir aucune information sur les vraies explications.

La deuxième option (ii) nous permet de résoudre ce problème et d'évaluer le système dans un contexte où les explications n'ont pas été construites manuellement. Néanmoins, le fait de conduire un processus d'instanciation automatique implique que le système ne pourrait pas toujours trouver des explications, parce que tout simplement il n'en existe pas dans le graphe sémantique. De plus, nous ne pourrions pas utiliser dans ce cas les métriques classiques d'évaluation des systèmes question réponse (précision, rappel, F-score, etc.), étant donné que celles-ci nécessitent de connaître à l'avance les bonnes réponses que le système doit retourner.

Après réflexion, nous avons choisi d'adopter la deuxième approche (ii), parce que nous souhaitons pour cette deuxième expérimentation connaître la performance réelle de notre plateforme dans un contexte où nous n'avons aucune influence sur les données.

4.4.1 Preuve de concept : données simulées et exécution manuelle

Comme expliqué dans l'introduction de ce chapitre, cette expérimentation a pour objectif de valider les résultats retournés par le processus d'explication des événements (contribution 2., section 1.3.2). Pour ce faire, nous nous focalisons sur l'explication de l'évènement *haut niveau de CO2* dans un parking connecté. Cet événement se déclenche quand un niveau élevé de CO2 est détecté (un capteur de CO2 détecte la concentration de CO2 et déclenche une alerte si cette concentration est supérieure à 1000 ppm). Le protocole expérimental est détaillé dans la section 4.4.1.1. Nous présentons et discutons les résultats dans la section 4.4.1.2.

Capteur	Nombre total observations
CO2	140 000
Température	140 000
Humidité	140 000
Présence voiture	720 000
Fumé	140 000
Mouvement	140 000
Total	1 420 000

TABLE 4.1 – Observations collectées par les 6 types de capteurs

4.4.1.1 Protocole expérimental

Dans cette section, nous présentons le jeu de données utilisé tout au long de la preuve de concept pour l'évaluation des résultats retournés par le processus d'explication des événements. Ensuite, nous présentons les métriques utilisées et nous détaillons la mise en œuvre.

Jeu de données

Nous présentons dans cette section le jeu de données de capteurs et le corpus documentaire utilisés dans cette preuve de concept. Ensuite, nous expliquons comment nous avons procédé pour définir et simuler l'événement haut niveau de CO2. Enfin, nous décrivons le choix des ontologies utilisées pour la modélisation des données du réseau de capteurs, du corpus documentaire et des événements.

— Données du réseau de capteurs

Les données de capteurs ont été simulées automatiquement à l'aide du module BLE de l'API nRF5SDK v11.0.0. Nous avons simulé les données de 78 capteurs liés au parking connecté (8 capteurs de CO2, 8 capteurs d'humidité, 8 capteurs de température, 40 détecteurs de présence de véhicules, 8 détecteurs de fumée et 8 détecteurs de mouvement). Les données simulées couvrent une durée de 5 heures. Chaque capteur produit une valeur toutes les 10 secondes. Un total de 1 420 000 observations ont été simulées. Le tableau 4.1 résume le nombre total des observations collectées par les six types de capteurs.

— Corpus de documents

Le corpus documentaire utilisé dans cette preuve de concept comporte un total de 20 fiches techniques de voitures et 15 autres pour les différents équipements de l'environnement (par exemple, le système de ventilation, le système de climatisation, le système d'éclairage, etc.). Ces documents ont été collectés sur internet.

— Définition et simulation de l'événement haut niveau de CO2

Pour définir l'événement haut niveau de CO2 dans l'ontologie réseau de capteurs,

nous avons procédé manuellement en utilisant l'éditeur Protégé⁹. Ensuite, nous avons simulé le déclenchement de cet événement en utilisant le module BLE de l'API nRF5SDK v11.0.0 mentionné plus haut.

— Ontologies de domaine

Pour modéliser les données du réseau de capteurs et les événements, nous avons utilisé l'ontologie HSSN-étendue (sections 2.2.2 et 2.5). Pour modéliser le corpus de documents, nous avons utilisé l'ontologie automobile COSAD [205] (Core Ontologies for Safe Autonomous Driving) et l'ontologie bâtiment SAREF4BLDG [117]. Nous avons choisi ces deux ontologies étant donné qu'elles couvrent l'ensemble des concepts et des relations sémantiques nécessaires à la représentation du parking connecté.

Après avoir détaillé le jeu de données utilisé dans le cadre de cette preuve de concept pour l'évaluation des résultats retournés par le processus d'explication des événements, nous présentons dans la section suivante les métriques d'évaluation et la mise en œuvre.

Métriques et mise en œuvre

Dans cette section nous présentons les métriques utilisées pour évaluer des résultats retournés par le processus d'explication des événements. Ensuite, nous détaillons la mise en œuvre.

— Métriques et critères d'évaluations

Nous conduisons des évaluations quantitatives et qualitatives des interconnexions sémantiques entre les données de capteurs et celles du corpus documentaire. En ce qui concerne l'évaluation quantitative, nous calculons le nombre total des interconnexions 5W1H construites puis filtrées dans chaque étape du processus d'explication d'événements. Quant à l'évaluation qualitative, nous utilisons les trois métriques *précision*, *rappel* et *F-score* communément utilisées dans la littérature pour l'évaluation des systèmes question-réponse [144].

— Mise en œuvre

Étant donné que cette preuve de concept a été menée avant le développement de la plateforme ISEEapp, l'exécution du processus d'explication des événements ainsi que le calcul des scores ont été réalisés manuellement.

Après avoir détaillé le protocole expérimental de la preuve de concept, nous présentons et discutons les résultats dans la section suivante.

4.4.1.2 Résultats

Dans cette section nous présentons et discutons les résultats de l'évaluation de la preuve de concept. Nous détaillons tout d'abord les résultats de l'évaluation quantitative. Ensuite, nous présentons les résultats de l'évaluation qualitative.

— Résultats de l'évaluation quantitative

9. <https://protege.stanford.edu/>

Pour évaluer la capacité du système à construire des interconnexions sémantiques entre les deux sources de données, nous avons résumé dans le tableau 4.2 le nombre total des interconnexions conceptuelles, les instances filtrées et les tuples Why_c produits par l'étape 1 du processus ISEE. Le tableau 4.3 présente le nombre total des interconnexions au niveau des instances et le nombre total des tuples Why_i produits par l'étape 2 du processus ISEE. Le tableau 4.4 présente le résultat produit par l'étape 3 du processus ISEE.

L'étape 1 du processus ISEE (tableau 4.2) établit un premier filtrage et réduit le graphe sémantique à un total de 26 concepts (13 concepts c_{What} reliés à l'événement par des connexions rc_{What} , 6 concepts c_{Who} reliés à l'événement par des connexions rc_{Who} , 3 concepts c_{When} reliés à l'événement par des connexions rc_{When} et 4 concepts c_{Where} reliés à l'événement par des connexions rc_{Where}). Ces 26 concepts relient à leur tour 284 instances candidates dans les ontologies HSSN-étendue, automobile et bâtiment. Enfin, l'étape 1 du processus ISEE a permis de construire 13 tuples Why_c non classés.

L'étape 2 du processus ISEE (tableau 4.3) réduit le graphe sémantique à un total de 35 instances (reliées à l'événement par différentes connexions sémantiques). Nous pouvons constater que le nombre d'instances potentiellement liées à l'événement est passé de 284 dans l'étape 1 à 35 dans l'étape 2. Cette baisse est due à l'exploitation des données de déclenchement de l'événement pour filtrer les instances (par exemple, l'heure et le lieu du déclenchement de l'événement). Par ailleurs, l'importance du niveau d'urgence est également apparente ici. Le processus de filtrage de l'étape 1 à l'étape 2 requiert un temps de traitement important, ce qui justifie la nécessité de niveaux d'urgence différents. Enfin, selon le niveau d'urgence de l'évènement l'étape 3 analyse les tuples construits par l'étape 1 (tuples Why_c) ou l'étape 2 (tuples Why_i) et calcule leurs scores de pertinence.

Les cinq tuples Why_i les mieux classés sont constitués de quatre instances de voitures (BMW G11 7, Infiniti Q60, VOLKSWAGEN Polo Match, Renault TCe 120) et d'une instance de système de ventilation (SAREF4BLDG:Fan-Webasto-5000). Comme déjà expliqué dans la section 4.4.1.1, les quatre voitures sont très polluantes (un niveau moyen élevé d'émissions de CO₂), tandis que le ventilateur a fait l'objet de plusieurs entretiens techniques au cours des deux derniers mois.

— Résultat de l'évaluation qualitative

Pour évaluer la qualité des interconnexions ISEE nous avons utilisé les métriques suivantes : **Précision (P)**, **Rappel (R)** et **F1-score** pour les étapes 1 et 2, et **Précision@n (P@n)** pour l'étape 3, étant donné que l'étape 3 retourne une liste ordonnée. Par ailleurs, nous nous sommes principalement concentrés sur l'évaluation des candidats de type What (c_{what} et i_{what}) et Who (c_{who} et i_{who}), car ce sont les parties qui nécessitent le traitement le plus complexe dans le système ISEE. Les résultats de l'évaluation sont présentés dans le tableau 4.5 (étapes 1 et 2) et le tableau 4.6 (étape 3). Dans l'étape 1 (tableau 4.5), la précision et le rappel sont presque les mêmes, puisque le nombre et la pertinence des concepts

Type du résultat	Nombre total
Connexions rc_{What}	13
Connexions rc_{Who}	6
Connexions rc_{When}	3
Connexions rc_{Where}	4
Instances de c_{What}	50
Instances de c_{Who}	63
Instances de c_{Where}	6
Instances de c_{When}	165
Tuples Why_c	6 tuples non classés
Total	26 concepts 284 instances 6 tuples Why_c non classés

TABLE 4.2 – Résultat de l'étape 1 du processus ISEE

Type du résultat	Nombre total
Connexions ri_{What}	13
Connexions ri_{Who}	11
Connexions ri_{Where}	1
Connexions ri_{When}	10
Tuples Why_i	11 tuples non classés
Total	19 concepts 35 instances 11 tuples Why_i non classés

TABLE 4.3 – Résultat de l'étape 2 du processus ISEE

Niveau d'urgence	Résultat étape 3	Exemples
Moyen	6 tuples Why_c classés	COSAD:Car, SAREF4BLDG:Fan, SAREF4BLDG:FireSuppressionTerminal, etc.
Faible	11 tuples Why_i classés	COSAD:BMW-G11-7, COSAD:Infiniti-Q60, COSAD:Renault-TCe-120, etc.

TABLE 4.4 – Résultat de l'étape 3 du processus ISEE

sélectionnés dépendent directement de la mesure de similarité sémantique (par exemple, les concepts `FireSuppressionTerminal` et `LowEmissionsCO2` sont des faux positifs qui ont été identifiés respectivement comme des candidats c_{what} et c_{who} puisque les concepts `Fire` et `Smoke` et les concepts `CO2` et `EmissionsCO2` sont sémantiquement proches). À l'étape 2 (tableau 4.5), le F1-score s'est amélioré, ce qui est dû au fait que les instances i_{what} et i_{who} sont sélectionnées en fonction des données exactes de déclenchement de l'événement (capteur, heure et lieu du déclenchement de l'événement), ce qui permet, dans la plupart des cas, de sélectionner les bons candidats. Cependant, les instances des concepts pertinents qui n'ont pas été sélectionnées à l'étape 1 sont ignorées par le processus de sélection. Par ailleurs, bien que l'instance de voiture '2018_Hyundai_H1_II_Cargo' ait un niveau d'émission de CO2 élevé, elle n'a pas été identifiée comme un candidat *Who*, car elle a quitté le bloc C deux minutes avant le déclenchement de l'événement. Une amélioration du processus de filtrage en fonction du type d'événement pourrait donc améliorer la qualité du processus de sélection des instances

à l'étape 2. Dans l'étape 3 (tableau 4.6), nous avons remarqué que la $P@n$ diminue lorsque la valeur de n augmente. Ceci est dû au fait que quelques tuples ont été mal classés. Néanmoins, le score pour les trois premières réponses est de 1, ce qui signifie que les bonnes réponses ont été retournées en premier. Nous remarquons également que les résultats pour le niveau d'urgence moyen sont généralement moins bons que ceux du niveau d'urgence faible, ceci est sûrement dû au fait que dans le cas du niveau d'urgence faible, le processus ISEE dispose de beaucoup plus d'informations pour construire et classer les tuples.

Attributs tuples Why	Tuples Why _c (Étape 1)		Tuples Why _i (Étape 2)	
	Concepts	Concepts	Instances	Instances
	c_{what}	c_{who}	i_{what}	i_{who}
P	0.66	0.92	0.76	1
R	0.83	0.92	0.81	0.84
F1-score	0.73	0.92	0.78	0.91

TABLE 4.5 – Évaluation des étapes 1 et 2 du système ISEE sur la base des trois paramètres d'évaluation suivants : précision, rappel et F1-score.

Précision@n (P@n)	P@3	P@5	p@8
Tuples Why _c	1	0.6	0.37
Tuples Why _i	1	0.8	0.75

TABLE 4.6 – Évaluation de l'étape 3 du processus ISEE en utilisant la métrique Precision@n

Pour résumer, l'évaluation quantitative a démontré que le système ISEE construit effectivement des interconnexions sémantiques entre le réseau de capteurs et les données du corpus documentaire. Par ailleurs, l'évolution du nombre de relations entre les étapes 1 et 2 dans le tableau 4.5 a démontré l'importance du choix du niveau d'urgence. Quant à l'évaluation qualitative, celle-ci a démontré que le système ISEE retourne effectivement des explications cohérentes aux événements déclenchés dans les environnements connectés (un score $P@3$ de 1 pour les deux niveaux d'urgence faible et moyen). Ceci montre l'efficacité de notre processus de filtrage.

Maintenant que nous avons détaillé la preuve de concept pour l'évaluation des résultats retournés par le processus d'explication des événements, nous passons dans la section suivante à l'expérimentation sur la plateforme ISEEapp. Cette expérimentation évalue également les résultats retournés par le processus d'explication des événements mais sur un jeu de données plus important, avec des données de capteurs réelles.

4.4.2 Expérimentation sur la plateforme ISEEapp : Données réelles et exécution automatique

Cette expérimentation a pour objectif d'évaluer les résultats du processus d'explication d'événements (contribution 2., section 1.3.2) sur un jeu de données plus large en utilisant la plateforme ISEEapp.

Nous nous focalisons dans cette expérimentation sur l'explication de l'évènement *gaspillage de lumière* dans un grand bâtiment de recherche. Cet évènement se déclenche lorsque des niveaux élevés successifs de luminosité sont détectés la nuit après les heures de travail dans le bâtiment. Le protocole expérimental est détaillé dans la section 4.4.2.1. Nous présentons et discutons les résultats respectivement dans la section 4.4.2.2.

4.4.2.1 Protocole expérimental

Dans cette section, nous présentons le jeu de données utilisé tout au long de l'expérimentation pour l'évaluation du processus d'explication d'événements. Ensuite, nous présentons les métriques utilisées et nous détaillons la mise en œuvre.

Jeu de données

Nous distinguons dans cette expérimentation deux jeux de données, le jeu de données de capteurs et le corpus documentaire. Nous expliquons par la suite comment nous avons procédé pour définir et détecter l'évènement *gaspillage de lumière*. Enfin, nous décrivons le choix des ontologies utilisées pour la modélisation des données du réseau de capteurs, du corpus documentaire et des événements.

■ **Données du réseau de capteurs**

Nous avons utilisé le jeu de données *Keti*¹⁰. Ce jeu de données est issu du bâtiment de recherche *Sutardja Dai Hall* de l'université de Californie. Les données de capteurs proviennent de 51 pièces réparties sur 4 étages. Chaque pièce comprend 5 types de mesures : la concentration de CO₂, le niveau d'humidité, la température ambiante, la luminosité et les données du capteur de mouvement PIR (Passive Infrared sensor). Ces données sont recueillies sur une période d'une semaine, du vendredi 23 août 2013 au samedi 31 août 2013. Le capteur de mouvement PIR est échantillonné une fois toutes les 10 secondes. Les autres capteurs sont échantillonnés une fois toutes les 5 secondes. Chaque fichier du jeu de données contient les horodatages et les relevés réels du capteur. Ces données sont utilisées par la suite pour instancier l'ontologie HSSN [116]. Le tableau 4.7 résume le nombre total d'observations collectées par les 5 types de capteurs.

■ **Corpus de documents**

Le corpus de document est composé de 150 fiches techniques reliées aux différents équipements installés dans le bâtiment de recherche (système d'éclairage, système de ventilation, ordinateurs, projecteurs, capteurs, etc.) et 130 fiches d'information des employés pour les différents types de poste (enseignant chercheur,

10. <https://www.kaggle.com/ranakrc/smart-building-system>

Capteur	Nombre total observations
Température	6 571 556
Humidité	6 571 516
CO2	6 574 059
Luminosité	6 571 514
Mouvement	3 594 004
Total	29 882 649

TABLE 4.7 – Observations collectées par les 5 types de capteurs

doctorant, ingénieur, etc.). La figure 4.3 représente une capture de la fiche d'information de l'employé *Philippe de la Charrier*. Le tableau 4.8 présente le nombre de fiches générées pour chaque type de poste.

Employee Information Sheet

Personal Information :

Philippe de la Charrier

E-mail: umahe@gmail.com Cell Phone: +33 (0)6 45 87 80 42

Address: 76, boulevard Torres, 61927 Marques, France Birth Date: 1959-04-04

Marital Status:

Job Information

Title	Employee ID	Star Date
laboratory director	C7X122	2005-07-06

Department	Assigned Office	Gross Annual Salary
IT	Office717	47860

Projects

Name	Start Date	End Date
MONT4	2018-10-13	2019-12-15
E2S	2016-11-18	-
BIS2	2018-02-12	2019-12-08

FIGURE 4.3 – Fiche d'information de l'employé Philippe de la Charrier

■ **Définition et détection de l'événement gaspillage de lumière**

Pour définir l'événement gaspillage de lumière dans l'ontologie réseau de capteurs, nous avons procédé manuellement en utilisant l'éditeur Protégé¹¹. Ensuite, pour détecter l'événement *gaspillage de lumière* dans le jeu de données Keti, nous nous sommes basés sur les données de deux capteurs : le capteur de luminosité et le capteur de mouvement. En effet, nous avons constaté en comparant les

11. <https://protege.stanford.edu/>

Poste	Nombre de fiches générées
Enseignant chercheur	40
Doctorant	30
Ingénieur	15
Stagiaire	8
Directeur de laboratoire	2
Responsable ressource humaine	5
Responsable logistique	5
technicien	15
Femme de ménage	5
Vigile	5
total	130

TABLE 4.8 – Nombre de fiches d’information générées par poste

données de capteur de lumière et de mouvement des mêmes salles, qu’il y a des cas où le niveau de luminosité est élevé durant la nuit (au dessus de 50 lux) simultanément avec des détections de mouvement. Par conséquent, pour s’assurer que l’événement *gaspillage de lumière* ne soit détecté que quand il n’y a personne dans la salle, nous avons ajouté la condition d’absence de mouvement détecté. Pour résumer, l’événement *gaspillage de lumière* est détecté quand les quatre conditions suivantes sont simultanément vérifiées :

- (i) un niveau de luminosité dépassant les 50 lux
- (ii) absence de mouvement
- (iii) les deux conditions (i) et (ii) doivent être vérifiées la nuit entre 20h30 et 5h.
- (iv) les trois conditions (i),(ii) et (iii) doivent être vérifiées durant un intervalle de temps qui dépasse 10 minutes.

En se basant sur ces conditions, nous avons implémenté une fonction python pour détecter le déclenchement de l’événement *gaspillage de lumière* dans le jeu de données Keti. Nous avons détecté au total 20 instances.

■ Ontologies de domaines

Pour modéliser les données du réseau de capteurs et les évènements nous avons utilisé l’ontologie HSSN-étendue. Ensuite, pour modéliser les données du corpus de documents, nous avons choisi les deux ontologies **SAREF4BLDG** [117] et **EOSK** [195] pour modéliser respectivement les données du domaine de bâtiment et des ressources humaines. Nous avons choisi ces deux ontologies parce qu’elles couvrent l’ensemble des concepts et des relations sémantiques nécessaires à la représentation de l’environnement connecté (le bâtiment de recherche). De plus, elles ont toutes les deux des structures simples et facile à manipuler.

Métriques et mise en œuvre

Dans cette section nous présentons les métriques utilisées pour évaluer le processus d'explication d'événements. Ensuite, nous détaillons la mise en œuvre de l'expérimentation.

— **Métriques et critères d'évaluations**

Comme nous l'avons déjà expliqué dans l'introduction de la section 4.4.2, étant donné que nous ne savons pas les vraies raisons derrière le déclenchement des instances de l'événement gaspillage de lumière, nous ne pouvons pas utiliser les métriques classiques d'évaluation des systèmes question-réponse (précision, rappel, f-score). Nous proposons donc d'utiliser deux métriques : **la complétude** et **la cohérence**.

- **La complétude** : cette métrique évalue le degré de complétude de l'explication retournée à l'utilisateur.

$$Comp = \frac{\alpha Cp_{what} + \beta Cp_{who} + \gamma Cp_{when} + \delta Cp_{where} + \theta Cp_{how}}{\alpha + \beta + \gamma + \delta + \theta}$$

— $\alpha, \beta, \gamma, \delta,$ and θ sont des coefficients de pondération

— $Cp_{what}, Cp_{who}, Cp_{when}, Cp_{where}$ et Cp_{how} font respectivement référence à la complétude des champs $c_{what}, c_{who}, c_{when}, c_{where},$ et c_{how} si le degrés d'explication demandé est standard ou bien $i_{what}, i_{who}, i_{when}, i_{where},$ et i_{how} s'il est d'une urgence élevée. La complétude d'un champ est égale à 1 si celui-ci est fourni dans l'explication et à 0 s'il ne l'est pas.

- **La cohérence** : c'est une mesure que l'expert donne à une réponse. La cohérence est égale à 1, si l'expert juge la réponse cohérente et 0 sinon. Par exemple, en supposant qu'un événement *gaspillage de lumière* se déclenche dans un bureau, une explication avec le champ *Who* (l'élément de l'environnement responsable du déclenchement) qui désigne le concept *extracteur d'air* est une réponse non-cohérente. La cohérence est donc égale à 0.

Nous avons testé les deux métriques de complétude et de cohérence lorsque nous avons conduit la preuve de concept (Section 4.4.1). Le jeu de données étant entièrement construit manuellement dans la preuve de concept, nous connaissions à l'avance les bonnes réponses. Nous avons constaté que la combinaison des deux métriques (complétude et cohérence) donnait des résultats similaire à 86% de ceux obtenus avec la métrique de précision.

— **Mise en œuvre**

Cette expérimentation a été exécutée sur la plateforme ISEEapp (section 4.2) pour obtenir les explications des 20 instances de l'événement gaspillage de lumière. Nous calculons la complétude et la cohérence des explications retournées par le système pour chaque instance de l'événement *gaspillage de lumière* et pour les deux niveaux d'explication *standard* et *complet*. Ensuite, nous calculons la moyenne des résultats obtenus. Nous nous sommes focalisés sur les niveaux *standard* et *complet* des explications, parce que ces deux niveaux nécessitent

Niveau de l'explication	Complétude min	Complétude max	Complétude moyenne @3	Complétude moyenne @10	Cohérence moyenne @3	Cohérence moyenne @10
Standard	0.60	1	1	0.93	1	0.80
Complet	0.40	1	0.82	0.64	0.85	0.71

TABLE 4.9 – Évaluation de la complétude et de la cohérence

une recherche et une construction de réponse à travers le graphe sémantique. Le calcul de la complétude est établi par une fonction python. La cohérence est, comme expliqué dans la section 4.4.2.1, mesurée par l'expert.

4.4.2.2 Résultats

Le tableau 4.9 représente les résultats de l'évaluation de la complétude et de la cohérence des vingt instances de l'événement *gaspillage de lumière* détectées dans le jeu données Keti (Section 4.4.2.1).

Concernant les résultats de l'évaluation de la complétude, les colonnes 2 à 5 du tableau 4.9 présentent respectivement la complétude minimale, la complétude maximale, la complétude moyenne des explications retournées dans les trois premières places et la complétude moyenne des explications retournées dans les dix premières places. Nous constatons, en comparant les colonnes 4 et 5 que les réponses bien classées par le système (i.e., retournée aux trois premières places) ont une meilleure moyenne de complétude. Nous remarquons également que la complétude du niveau *complet* (Ligne 3, tableau 4.9) est inférieure à celle du niveau *standard* (Ligne 2, tableau 4.9). Ceci est peut-être dû au fait que nous pouvons parfois avoir des concepts liés aux éléments de l'environnement mais que les instances de ces derniers ne sont pas renseignées. Par exemple, pour le concept *projecteur*, nous pouvons savoir dans quelle pièce il est placé, mais nous ne pouvons pas savoir quand il est allumé et nous n'avons donc pas d'information exacte sur les périodes de son fonctionnement.

Concernant les résultats de l'évaluation de la cohérence, les colonnes 6 et 7 du tableau 4.9 présentent respectivement la cohérence moyenne des explications retournées aux trois premières places et la cohérence moyenne des explications retournées en dix premières places. Comme pour la complétude, nous constatons en comparant les deux colonnes que les réponses bien classées par le système ont une meilleure moyenne de cohérence. La baisse de niveau de cohérence en passant au niveau d'explication complet, peut être expliquée par le fait qu'un seul candidat non-cohérent au niveau conceptuel peut en générer plusieurs au niveau des instances (un concept peut avoir plusieurs instances). La présence de candidats non-cohérents est généralement due soit aux erreurs au niveau du processus d'instanciation soit aux alignements erronés. Finalement, il faut savoir que parmi les vingt instances de l'événement *gaspillage de lumière*, le système n'a pu retrouver d'explications pour trois d'entre elles. Ceci peut-

être expliqué par le fait que l’instanciation des éléments de l’environnement a été faite de manière aléatoire (Section 4.4.2.1). Par conséquent, parfois le jeu de données ne comporte aucun élément pouvant expliquer le déclenchement de l’événement.

Pour résumer, l’évaluation de la complétude et de la cohérence nous ont permis de valider notre processus d’interconnexion et de filtrage (une complétude et une cohérence au-dessus de 0.82 pour les réponses retournées en 3 premières places). Les résultats ont également confirmé une deuxième fois l’importance du choix de l’outil d’alignement mais également celle du processus d’instanciation.

4.5 Évaluation de l’interface utilisateur pour l’explication d’événements

Dans cette section, nous évaluons l’interface utilisateur de la plateforme ISEEapp (contributions (i) et (ii)). Nous souhaitons savoir si les réponses retournées sont suffisamment claires et faciles à comprendre par les utilisateurs de l’environnement connecté. Nous conduisons une enquête auprès d’un groupe d’utilisateurs potentiels. Ensuite, nous calculons les moyennes des niveaux de compréhension. Nous détaillons le protocole expérimental dans la section 4.5.1. Nous présentons et discutons les résultats dans la section 4.5.2.

4.5.1 Protocole expérimental

L’enquête pour l’évaluation de l’interface utilisateur ISEEapp comprend deux exemples différents de l’événement *gaspillage de lumière* (section 4.4.2.1). Nous détaillons dans ce qui suit, le jeu de données utilisé pour générer ces événements (section 4.5.1.1). Ensuite, nous présentons la métrique utilisée et nous détaillons la mise en œuvre (section 4.5.1.2).

4.5.1.1 Jeu de données

Pour capturer les deux exemples de l’événement *gaspillage de lumière*, nous avons alimenté la plateforme ISEEapp avec le même jeu de données de la section 4.4.2.1.

4.5.1.2 Métriques et mise en œuvre

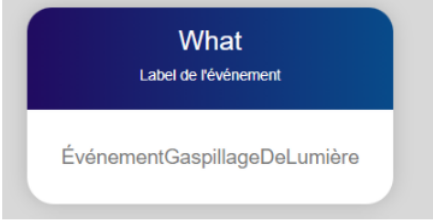
Afin d’évaluer la facilité de compréhension des explications retournées par le système, nous avons créé un questionnaire¹² avec deux exemples différents de l’événement *gaspillage de lumières* et leurs explications. Nous commençons tout d’abord par introduire le concept général du système et son utilité. Nous présentons la structure générale des réponses retournées et nous expliquons la signification de chaque champ 5W1H. Ensuite, nous montrons des captures d’écrans des réponses retournées pour les deux

12. Le questionnaire est disponible sur le lien suivant : <https://forms.gle/ftC5qAv4EzQNv6cU6>

Pour rappel

- What = Quel est l'évènement déclenché.
- Who = Qu'est-ce qui a provoqué le déclenchement de l'évènement.
- When = Quand l'évènement s'est-il déclenché.
- Where = Où l'évènement s'est-il déclenché.
- How = Comment l'évènement s'est-il déclenché.
- Why = Quelles sont les raisons potentielles du déclenchement de l'évènement.

Est-ce que vous comprenez la signification du critère "What" de ce visuel ? *



Oui, je comprends parfaitement ce visuel
 Je comprends partiellement ce visuel
 Non, je ne comprends pas ce visuel

FIGURE 4.4 – Capture d'écran du questionnaire d'évaluation de facilité de compréhension

exemples de l'évènement *gaspillage de lumière*. Après, nous posons la question 'est ce que vous comprenez la signification du champ?' pour chaque élément 5W1H. L'utilisateur peut répondre aux questions par 'Oui je comprend parfaitement', 'je comprends partiellement' ou 'je ne comprend pas'. Une capture d'écran d'une question sur le champ *What* est présentée dans la figure 4.4. Pour le premier exemple de l'évènement *gaspillage de lumière*, nous ne détaillons pas la question par rapport aux champs contenus dans la réponse à la question *Why* (i.e., *Why*→*What*, *Why*→*Who*, *Why*→*When*, *Why*→*Where*, *Why*→*How*), l'idée étant ici de savoir si un *Why* de façon générale est compréhensible ou non. Par ailleurs, pour des raisons de temps, nous ne pouvions pas détailler les questions sur le champ *Why* pour les deux évènements. Nous avons conduit l'enquête auprès de deux classes (au totale 62 étudiants) : première année informatique (25) et première année génie industriel et maintenance GIM (37 étudiants). Nous avons choisi ces deux classes respectivement pour leur aisance avec les outils de recherche d'information et pour leur familiarisation avec le domaine du bâtiment. Les résultats de l'enquête sont présentés et discutés dans la section 4.5.2.

4.5.2 Résultat

Les figures 4.5 et 4.6 représentent les résultats de l'évaluation de la facilité de compréhension pour les deux exemples de l'évènement *gaspillage de lumière*. Les histogrammes en gris dans les deux figures mesurent la moyenne pour les deux classes. Pour le deuxième exemple (figure 4.6), le champ *Why*→*When* n'a pas été retourné par le

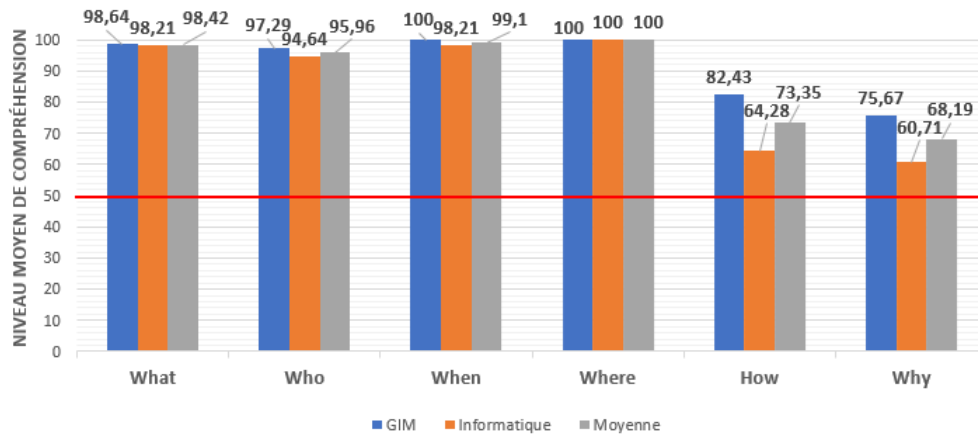


FIGURE 4.5 – Évaluation de la facilité de compréhension sur le 1^{er} exemple de l'événement gaspillage de lumière

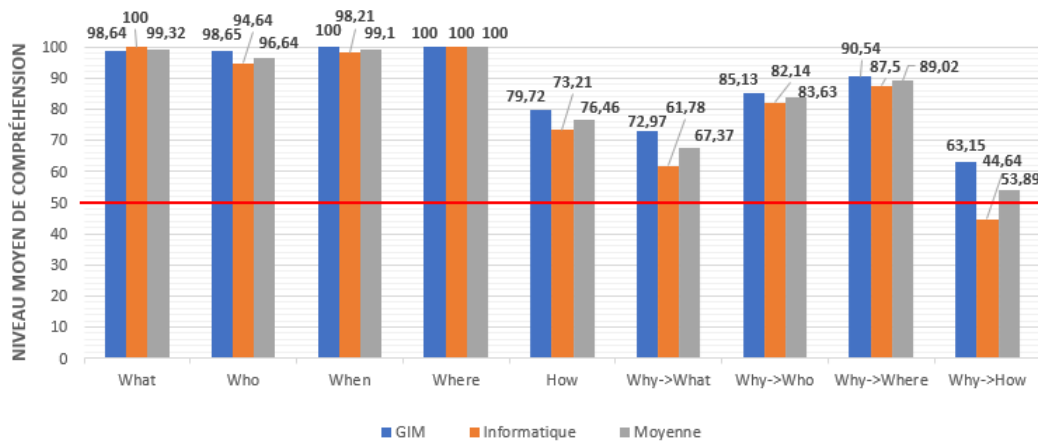


FIGURE 4.6 – Évaluation de la facilité de compréhension sur le 2^{ème} exemple de l'événement gaspillage de lumière

système, c'est pour cela qu'il n'est pas représenté dans la figure. Le but ici est de savoir si les résultats retournés par le système sont toujours compréhensibles même si un champ n'est pas renseigné. Concernant les quatre champs *What*, *Who*, *When* et *Where*, nous constatons que les résultats sont tous supérieurs à 94% dans les deux figures. Nous concluons qu'il n'y a pas d'ambiguïté par rapport à la compréhension de ces quatre champs. Les champs *How* et *Why* sont globalement un peu plus ambiguës, néanmoins, les moyennes restent au-dessus de 64% dans les deux figures. Concernant les 5 champs, nous détaillons la réponse à la question *Why* (*Why*→*What*, *Why*→*Who*, *Why*→*When*, *Why*→*Where* et *Why*→*How*, figure 4.6), le champ *Why*→*How* est celui qui pose le plus de difficultés surtout pour les étudiants d'informatique. Les histogrammes en gris dans les deux figures montrent que globalement, tous les résultats sont au-dessus de la barre de 50%. Les étudiants de la filière informatique ont eu plus de difficulté, ceci est peut-être

du au faite qu'ils ne sont pas aussi bien familiarisés avec les termes et les notions utilisés dans le domaine du bâtiment que les étudiants de la filière GIM.

Pour résumer, l'évaluation de la facilité de compréhension nous a permis de valider le modèle de structuration des explications des évènements (la moyenne de compréhension pour tous les champs est au-dessus de 50%). Elle nous a permis également d'identifier les champs ambigus du modèle. Les résultats de l'enquête ont démontré que les attributs *Why*→*What* et *Why*→*How* de la réponse à la question *Why*, sont ceux qui posent le plus de difficulté de compréhension aux utilisateurs, surtout pour ceux qui ne sont pas familiarisés avec le domaine du bâtiment. Les améliorations futures à établir sur notre modèle 5W1H pour la structuration des explications seront présentées dans la section perspectives de cette thèse (section 5.2).

Maintenant après avoir détaillé les différentes expérimentations conduites pour valider nos contributions, nous concluons ce chapitre dans la section suivante.

4.6 Conclusion

Ce chapitre présente une évaluation expérimentale des contributions du chapitre 3. Nous avons tout d'abord présenté l'architecture générale de la plateforme ISEEapp qui concrétise ces contributions sous forme d'une application Web dans la section 4.2. Ensuite, la section 4.3 présente une expérimentation qui évalue notre modèle de définition des événements (contribution 1.1., section 1.3.1). La section 4.4 inclut deux expérimentations qui évaluent la qualité des réponses retournées par le processus d'explication d'évènements (contribution 2., section 1.3.2). Enfin, la section 4.5 décrit une évaluation de l'interface utilisateur de la plateforme ISEEapp (contribution 1. et 2., section 1.3).

La première expérimentation (section 4.3), avait pour objectif d'évaluer notre modèle de définition des événements (contribution 1.1., section 1.3.1). Elle consistait à mesurer l'évolution de la connectivité du graphe sémantique selon la diversité des évènements définis dans ce dernier. Elle nous a permis de valider notre modèle de définition des événements et d'affirmer que ce dernier permet réellement de favoriser le rapprochement des données du réseau de capteurs et les données du corpus documentaire (une meilleure évolution de la connectivité du graphe sémantique, dans le cas de l'utilisation de notre modèle comparé à deux autres modèles pour la définition d'évènements).

Concernant les deux expérimentations de la section 4.4, la preuve de concept (section 4.4.1) avait pour but de conduire une évaluation quantitative et qualitative des interconnexions sémantiques (contribution 2., section 1.3.2). Elle nous a permis de tirer la conclusion suivante :

- L'évaluation quantitative (Section 4.4.1.2) a démontré que le système ISEE construit effectivement des interconnexions sémantiques entre les données réseau de capteurs et celles du corpus documentaire (pour l'instance de l'évènement *haut*

niveau de CO2, 26 concepts ont été connectés, 35 instances ont été connectées et filtrées, 6 tuples Why_c et 11 tuples Why_i ont été construits);

- L'évaluation qualitative (Section 4.4.1.2) a démontré que le système ISEE retourne effectivement les explications aux événements déclenchés dans les environnements connectés (un score P@5 de 0.6 et P@3 de 1 pour les explications standards et un score P@5 de 0.8 et P@3 de 1 pour les explications complètes).

La deuxième expérimentation (section 4.4.2) avait pour objectif d'évaluer le processus d'explication d'événements (contribution 2., section 1.3.2) sur un jeu de données plus large en utilisant la plateforme ISEEapp. Nous avons proposé deux métriques originales : la complétude et la cohérence. Cette expérimentation nous a permis de valider notre processus d'interconnexion et de filtrage (une complétude et une cohérence au-dessus de 0.82 pour les réponses retournées dans les 3 premières places).

La dernière expérimentation (section 4.3), visait l'évaluation de l'interface utilisateur de la plateforme ISEEapp (contributions (i) et (ii)) à travers une enquête. Cette enquête consistait à mesurer la facilité de compréhension des différents champs 5W1H auprès des utilisateurs. Elle nous a permis de valider le processus de construction d'explication ainsi que notre modèle 5W1H pour la structuration des explications (la moyenne de compréhension pour tous les champs est au-dessus de 50%). Toutefois, des améliorations sont à prévoir sur les champs *Why*→*What* et *Why*→*How* identifiés comme les plus ambiguës pour les utilisateurs.

Chapitre 5

Conclusion

5.1 Récapitulatif

Cette thèse a pour objectif l'explication des événements détectés dans les environnements connectés, et plus précisément ceux qui se produisent dans des environnements disposant de systèmes d'information hybrides (un système d'information gérant le corpus de documents et un système d'information gérant les données du réseau de capteurs). Nous avons proposé un système intitulé ISEE (Information System for Event Explanation). ISEE est basé sur : (i) un modèle multidimensionnel pour la définition des événements dans les environnements connectés, (ii) un modèle 5W1H pour la structuration des explications d'événements, (iii) un processus pour l'interconnexion et le filtrage ciblé des concepts et des instances des ontologies de domaine et (iv) un processus pour le classement des explications. Dans ce qui suit, nous résumons ce que nous avons vu dans chaque chapitre de cette thèse.

Dans **le chapitre 1**, nous avons introduit les deux domaines d'activités liés à la thèse, à savoir, les environnements connectés et la digitalisation des entreprises. Nous avons tout d'abord donné un aperçu des raisons pour lesquelles ces deux domaines sont considérés comme des sujets d'intérêt de nos jours et nous avons présenté quelques statistiques récentes.

Nous avons détaillé l'objectif de cette thèse : l'explication des événements détectés dans les environnement hybrides. Ensuite, nous avons présenté deux scénarios qui illustrent les motivations de ce travail et les défis qui en découlent : (i) un événement gaspillage de lumière dans un bâtiment de recherche connecté et (ii) un événement haut niveau de CO₂ dans un parking connecté.

Les principales contributions de cette thèse ont été présentées ((i) modélisation des événements dans les environnements hybrides, (ii) interconnexion et filtrage ciblés des ontologies de domaine), les différentes étapes ont été expliquées et illustrées à travers le scénario de l'événement gaspillage de lumière. Enfin, nous avons listé les publications liées à cette thèse avant d'introduire les chapitres suivants.

Dans **le chapitre 2**, nous avons présenté l'état de l'art autour des trois axes de recherche liés à cette thèse : la recherche d'information classique et sémantique dans des environnements connectés, la recherche d'information classique et sémantique dans les corpus documentaires et l'interconnexion de données hétérogènes.

Nous avons passé en revue les différentes approches et techniques proposées. Ensuite, nous avons identifié les éléments qui semblent intéressants pour résoudre chacun des défis présentés dans le chapitre 1, à savoir (i) l'ontologie HSSN pour modéliser les environnements connectés ainsi que les données de capteurs; (ii) l'approche 5W1H pour guider le processus de recherche et structurer de façon simple et claire les explications; (iii) les techniques d'alignement d'ontologies pour rapprocher les ontologies de domaine au niveau conceptuel; et (iv) les techniques de liaison de données pour interconnecter les ontologies au niveau des instances.

Dans **le chapitre 3**, nous avons présenté notre proposition, le système ISEE pour l'explication des événements dans les environnements connectés. À notre connaissance, ISEE est le premier système dédié à l'explication des événements utilisant à la fois les données de réseau de capteurs et les données de corpus documentaires.

Le système ISEE s'appuie sur (i) notre modèle multidimensionnel pour la définition des événements dans les environnements connectés : ce modèle permet aux utilisateurs de définir les événements qu'ils souhaitent détecter selon différents axes de description (capteurs et documents), pour ainsi rapprocher les données du réseau de capteurs et du corpus documentaires par le biais de ces définitions; (ii) notre modèle 5W1H pour la structuration des explications d'événements : ce modèle s'inspire de l'approche 5W1H et l'adapte à notre contexte pour présenter aux utilisateurs des explications simples et faciles à comprendre; (iii) notre processus pour l'interconnexion et le filtrage ciblé des ontologies de domaine : ce processus a pour objectif d'analyser les données du graphe sémantique (l'ensemble des ontologies de domaine après leur instanciation et leur alignement) et de construire des explications aux événements déclenchés dans l'environnement en s'appuyant sur des interconnexions sensibles au contexte; et, enfin, (iv) notre processus pour le classement des explications : ce processus a pour objectif de classer par ordre de pertinence les explications construites par l'étape précédente en s'appuyant sur une métrique originale.

Dans **le chapitre 4**, nous avons tout d'abord présenté l'application Web ISEEapp qui met en œuvre les contributions du chapitre 3. Ensuite, nous nous sommes focalisés sur l'évaluation de ces contributions. Nous avons mené trois évaluations expérimentales : (i) l'évaluation du modèle de définition des événements (ii) l'évaluation des résultats du processus d'explication des événements et (iii) l'évaluation de l'interface utilisateur pour l'explication des événements.

La première expérimentation (i) visait à évaluer la capacité du modèle de définition des événements à rapprocher les données du réseau de capteurs et du corpus de documents. Pour ce faire, nous avons mesuré l'évolution de la connectivité du graphe sémantique selon la diversité des événements définis dans celui-ci. Le résultat de cette

expérimentation nous a permis de valider notre modèle de définition d'événement (une meilleure évolution de la connectivité du graphe sémantique dans le cas de l'utilisation de notre modèle comparé à deux autres modèles pour la définition des événements).

Pour évaluer les résultats du processus d'explication des événements (ii), nous avons mené deux expérimentations : (a) une preuve de concepts manuelle sur un petit jeu de données simulé et (b) une expérimentation sur la plateforme ISEEapp avec un jeu de données plus large et des données de capteurs réelles. La preuve de concept (a) avait pour but d'établir des évaluations quantitatives et qualitatives des interconnexions sémantiques entre les données du réseau de capteurs et celles du corpus documentaire en s'appuyant sur des métriques classiques (précision, rappel et F-score). L'expérimentation sur la plateforme ISEEapp (b) avait pour objectif d'évaluer les résultats du processus d'explication des événements en s'appuyant sur des métriques originales (complétude et cohérence). Les deux expérimentations (a) et (b) ont abouti à des résultats prometteurs qui nous motivent à poursuivre l'amélioration de notre proposition (une complétude et une cohérence au-dessus de 0.82 pour les réponses retournées dans les 3 premières places).

La dernière expérimentation (iii) avait pour but d'évaluer la facilité de compréhension des résultats retournés par la plateforme ISEEapp. Pour ce faire, nous avons mené une enquête auprès d'un groupe d'utilisateurs potentiels. Ensuite, nous avons calculé les moyennes des niveaux de compréhension. Le résultat de cette expérimentation nous a permis de valider le processus de construction des explications ainsi que notre modèle 5W1H pour la structuration des explications (la moyenne de compréhension pour tous les champs est au-dessus de 50%). Toutefois, des améliorations sont à prévoir sur quelques champs identifiés comme les plus ambiguës par les utilisateurs.

5.2 Perspectives

Plusieurs travaux futurs peuvent être menés pour améliorer plus les contributions de cette thèse. Dans ce qui suit, nous présentons trois perspectives principales qui se positionnent respectivement au niveau du processus d'intégration des connaissances de domaine ((A), figure 5.1), le processus ISEE ((B), figure 5.1) et des données de sortie du système ISEE ((C), figure 5.1).

Développement de la fonction d'instanciation des ontologies de domaine à partir du corpus documentaire ((A), figure 5.1)

Comme nous l'avons expliqué dans la section 4.2, nous n'avons pas encore développé la fonction qui instancie automatiquement les ontologies de domaine à partir du corpus documentaire. Cette fonction joue un rôle essentiel dans le système ISEE. Elle permet d'accéder à l'ensemble des informations contenues dans le corpus documentaire à travers les ontologies de domaine et ainsi exploiter ces données pour construire les explications. C'est donc la prochaine étape sur laquelle nous allons nous concentrer. Cette perspective concerne principalement le verrou scientifique du traitement

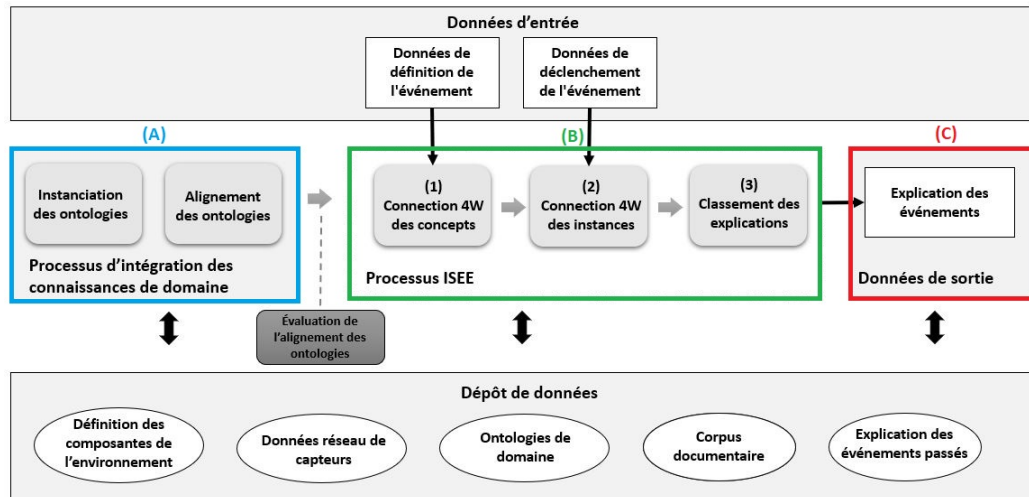


FIGURE 5.1 – Le système ISEE et les perspectives

automatique du langage naturel appliqué à des corpus de documents hétérogènes (des documents semi- ou non structurés couvrant plusieurs domaines d'application).

Les approches de la littérature pour l'instanciation automatique des ontologies à partir de texte sont généralement basées sur les techniques de traitement automatique des langages naturels [38, 113] associées avec des techniques d'apprentissage automatique [162, 165] et/ou des techniques d'extraction d'information (par exemple, la reconnaissance des entités nommées) [35]. L'approche présentée dans [38] nous paraît être un bon point de départ. Elle est générique (peut être appliquée dans n'importe quel domaine d'application) et peu coûteuse en termes de ressources. De plus, l'évaluation menée dans leurs articles montre de très bons résultats. Pour implémenter cette approche nous avons identifié les bibliothèques Python *SpaCy*¹ et *NLTK*² dédiées au traitement des langages naturels et à l'extraction d'information, ainsi que la bibliothèque *TensorFlow*³ spécialisée dans les algorithmes d'apprentissage automatique. Par ailleurs, l'outil *Gate*⁴ qui est très populaire et qui est dédié au traitement de texte, nous paraît également être une bonne piste à explorer. Nous avons identifié les deux modules *ONTOTEXT KIM*⁵ et *OWLLIM*⁶ de l'outil *Gate* qui sont spécifiquement dédiés aux ontologies et aux graphes sémantiques en général.

Prise en considération des événements complexes ((B), figure 5.1)

Pour l'instant, le système ISEE ne traite que les événements de type atomique. Les événements atomiques sont les événements les plus simples qui puissent se produire

1. <https://spacy.io/>
2. <https://www.nltk.org/>
3. <https://www.tensorflow.org/>
4. <https://gate.ac.uk/>
5. <https://rb.gy/bgyjzw>
6. <https://rb.gy/tl0n7h>

dans un système. Ils ne peuvent pas être décomposés en une entité plus petite (par exemple, *haut niveau de température*, *haut niveau de CO2*, etc.) [115]. Nous voulons étendre notre système ISEE pour gérer les événements complexes (appelés aussi événements composites). Les événements complexes sont les événements qui combinent des événements constitutifs. Ces derniers peuvent être atomiques et/ou complexes (par exemple, un incendie est une combinaison des deux événements *haut niveau de température* et *présence de fumée*) [115].

Pour rappel, dans notre modèle de définition d'événements, chaque événement est caractérisé par un espace à 4 dimensions appelé eSpace (déf. 13). L'eSpace se compose des dimensions Feature, Source, Time, et Location ainsi que des données de l'événement pour représenter les observations des capteurs qui ont permis de le détecter. Si nous restons sur l'exemple de l'événement incendie, notre modèle de définition d'événements permet de définir les eSpace des événements *haut niveau de température* (eSpace_{ht}) et *présence de fumée* (eSpace_{pf}) comme suit :

eSpace_{ht} : ⟨1, Feature_{ht}, Source_{ht}, Time_{ht}, Location_{ht}, I_{ht}⟩, tel que :

- **Feature_{ht}** : ⟨2, Temperature, {Heat, Warmth}, {hasTemperature}, {Temperature > 35}⟩
- **Source_{ht}** : ⟨3, TemperatureSensor, {AirConditioner}, {senses, makesObservation}, {}⟩
- **Time_{ht}** : ⟨4, TimeInterval, {TemporalEntity, TimeInstant}, {inDateTime, inTemporalPosition}, {}⟩
- **Location_{ht}** : ⟨5, Location, {Coordinate, Position, Floor, Room, SpatialArea}, {hasLocation}, {}⟩
- **I_{ht}** : {}

eSpace_{pf} : ⟨6, Feature_{pf}, Source_{pf}, Time_{pf}, Location_{pf}, I_{pf}⟩, tel que :

- **Feature_{pf}** : ⟨7, Smoke, {Fog, Pollution}, {producesSmoke}, {SmokePresence = True}⟩
- **Source_{pf}** : ⟨8, SmokeDetector, {AirConditioner, Stove}, {senses, makesObservation}, {}⟩
- **Time_{pf}** : ⟨9, TimeInterval, {TemporalEntity, TimeInstant}, {inDateTime, inTemporalPosition}, {}⟩
- **Location_{pf}** : ⟨10, Location, {Coordinate, Position, Floor, Room, SpatialArea}, {hasLocation}, {}⟩
- **I_{pf}** : {}

Maintenant, pour définir l'événement *incendie* (la combinaison des deux événements *haut niveau de température* et *présence de fumée*), il suffit d'agréger ces deux définitions. Pour cela, il faut vérifier que ces deux événements se déclenchent dans le même périmètre spatial et temporel. Par conséquent, il faut compléter la définition de l'événement *incendie* par la **condition de composition** suivante : les deux événements

haut niveau de température et *présence de fumée* doivent être déclenchés simultanément dans le même périmètre spatio-temporel.

Pour résumer, afin de traiter les événements complexes dans le système ISEE nous introduisons des ajustements dans notre modèle de définition des événements (section 3.4.1). Nous définissons un événement complexe comme (i) un ensemble d'événements simples ou complexes associés à (ii) une ou plusieurs conditions de composition de ces événements, tel que :

$$EC = \langle \bigcup_{i=0}^n e_i, \bigcup_{j=0}^m c_j \rangle \quad \forall i, j \in \mathbb{N}$$

Où :

- **EC** est l'acronyme d'événement complexe, e_i est un des événements composant l'événement *EC*, c_j est une condition de composition de ces événements.

Notez ici que les conditions de composition peuvent être elles-mêmes simples (par exemple, " e_i .Location $\Leftrightarrow e_j$.Location" pour indiquer que les deux événements e_i et e_j doivent avoir lieu dans un même périmètre spatial) ou complexes (plusieurs conditions simples agrégées avec des opérateurs logiques, par exemple, " e_i .Location $\Leftrightarrow e_j$.Location AND e_i .Time $\Leftrightarrow e_j$.Time" pour indiquer que les deux événements e_i et e_j doivent avoir lieu dans le même périmètre spatial et temporel). Bien entendu, la définition proposée doit être testée pour plusieurs événements complexes.

Une fois l'étape d'ajustement de notre modèle de définition d'événements terminée, l'étape suivante sera l'adaptation du processus ISEE (section 3.5.3). Ici, plusieurs questions se posent : Sachant que le processus ISEE dans sa version actuelle construit des pistes d'explications pour les événements atomiques, comment combiner ces pistes d'explications pour construire des explications à un événement complexe composé de ces événements ? Faut-il s'appuyer sur les conditions de composition c_j ? ou bien faut-il privilégier les pistes d'explications retournées par plusieurs événements composants l'événement complexe ? Pour répondre à ces questions, la première étape que nous souhaitons entreprendre consiste à collecter un grand nombre d'exemples d'événements complexes avec leurs explications respectives, puis tirer des observations à partir de ces exemples. Nous pensons qu'en suivant cette méthodologie, nous serons en mesure de voir le problème plus clairement et même de proposer de nouvelles idées.

Amélioration de la clarté et de la lisibilité des explications ((C), figure 5.1)

Pour lever l'ambiguïté sur certains champs des explications construites par le système et améliorer la facilité de compréhension de ces explications par les utilisateurs, nous envisageons les modifications suivantes :

- Accompagner chaque champ d'une ou deux phrases pour rappeler à l'utilisateur à chaque fois la signification du champ.
- Ne pas mettre l'IRI des concepts mais plutôt leurs labels pour éviter de perturber les utilisateurs qui n'ont aucune connaissance du web sémantique. Les utilisateurs avancés pourront obtenir l'IRI en double-cliquant sur le champ.

- Modifier la représentation du champ *Why→How*. Ce champ est représenté sous forme de triplets RDF (*< sujet, prédicat, objet >*). Nous modifions légèrement cette représentation, et nous formulons plutôt une phrase avec les trois éléments du triplet.
- Inverser la position des deux champs *Why→What* et *Why→Who*. Les champs *Why→Who* et *Why→What* représentent respectivement l'entité responsable du déclenchement de l'événement et l'élément qui explique pourquoi celle-ci a été identifiée comme responsable. Représenter ces deux champs dans cet ordre (entité responsable puis l'élément expliquant pourquoi) nous semble être plus logique et facile à comprendre pour les utilisateurs.

Annexe A

ISEEapp

A.1 Introduction

Dans cet annexe, nous présentons la plateforme ISEEapp qui concrétise le système ISEE sous forme d'une application Web.

La plateforme ISEEapp est constituée de deux couches, **une couche services** et **une couche accès aux ressources**. La couche services comprend six modules. Nous les organisons en trois catégories, à savoir (i) **intégration des connaissances de domaine** qui correspond aux différentes étapes du processus d'intégration des connaissances de domaine (intégration des connaissances de domaine, figure 3.1); (ii) **gestion des événements** qui permet à l'utilisateur de définir des événements, puis détecte les occurrences de ces événements dans les données de capteurs (données d'entrée, figure 3.1); (iii) **processus ISEE** qui englobe les différentes étapes du processus ISEE (processus ISEE, figure 3.1). La couche accès aux ressources comprend les fonctions python pour la lecture et l'écriture dans le dépôt de données.

Notez ici que lors du développement de la plateforme ISEEapp, nous nous sommes focalisés, dans un premier temps sur les modules qui constituent des contributions de cette thèse et sur ceux qui sont nécessaires au bon fonctionnement de la plateforme. Nous ne gérons pas pour l'instant l'aspect dynamique des données de capteurs (la collecte automatique des données à partir des capteurs installés dans l'environnement). Dans la version actuelle de la plateforme, nous supposons que nous disposons d'un jeu de données de capteurs comportant plusieurs occurrences d'événements. L'utilisateur peut ajouter des documents ou des données de capteurs à travers la plateforme.

La suite de cet annexe est organisée comme suit. Dans la section 4.2 nous présentons l'architecture générale de la plateforme ISEEapp et nous précisons les choix techniques du développement. Les sections A.2, A.3, A.4, A.5 et A.7 détaillent respectivement le développement des différents modules de la plateforme et présentent quelques captures d'écrans de la plateforme ISEEapp.

A.2 Instanciation des ontologies

L'objectif principal de ce module est d'instancier automatiquement les ontologies de domaine (ontologie de réseau de capteurs et les ontologies représentant la sémantique du corpus documentaire) à partir du dépôt de données. Pour l'instant, le processus d'instanciation n'est encore automatisé que pour les données de capteurs, l'instanciation des ontologies de domaine à partir du corpus documentaire se fait manuellement nous détaillons ces deux processus dans ce qui suit.

- **Instanciation de l'ontologie HSSN-étendue avec les données de capteurs** : pour instancier automatiquement l'ontologie HSSN-étendue avec les données de capteurs, nous avons implémenté une fonction python. Cette fonction prend en entrée :
 - **L'id du capteur** : cet argument identifie de manière unique le capteur (par exemple, 'CO2Sensor1').
 - **La localisation du capteur** : cet argument désigne l'emplacement du capteur dans l'environnement connecté (par exemple, 'Block C').
 - **La propriété capturée** : ce dernier indique quelle propriété physique est capturée par le capteur (par exemple 'CO2').
 - **Le fichier contenant les données capturées** : cet argument est renseigné à travers le chemin d'accès au fichier (par exemple, '.\Sensor_Data\BlockC\CO2.csv').

La fonction d'instanciation s'appuie sur le paquet *Owlready2*¹ précédemment cité dans la section 4.2. Pour instancier l'ontologie HSSN-étendue, la fonction fait une mise en correspondance directe entre les arguments fournis et les concepts de l'ontologie HSSN-étendue. Notez ici que lorsque de nouvelles données de capteurs sont ajoutées par l'utilisateur le processus d'instanciation est exécuté automatiquement par la plateforme pour que celles-ci intègrent le graphe sémantique.

- **Instanciation des ontologies de domaine à partir du corpus documentaire** : comme nous l'avons expliqué plus haut, dans la version actuelle de la plateforme l'instanciation des ontologies de domaine à partir du corpus documentaire se fait manuellement. Nous analysons les documents et nous les traduisons sous forme de données tabulaires qui sont ensuite utilisées pour instancier les ontologies de domaine. L'automatisation de ce processus est traitée dans la section perspectives de cette thèse (section 5.2).

Après avoir présenté le module d'instanciation des ontologies de domaine, la section suivante se consacre au module d'alignement des ontologies.

1. <https://owlready2.readthedocs.io/en/v0.35/index.html>

A.3 Alignement des ontologies

Ce module a pour objectif d'aligner les ontologies de domaine en s'appuyant sur des outils existants. La version actuelle de la plateforme intègre les deux outils **AgreementMakerLight**² et **TheAlignmentAPI**³. Nous avons choisi ces outils car ils sont très communément utilisés dans la littérature. De plus, ils sont performants et faciles à utiliser. La figure A.1 représente une capture d'écran d'un fichier d'alignement retourné par l'outil AgreementMakerLight pour une ontologie temporelle (time) et une ontologie ressources humaine (HR-Ontology). Nous pouvons voir une correspondance entre le concept *Instant* de l'ontologie *Time* et le concept *TimeInstant* de l'ontologie *HR-Ontology*. Cette correspondance possède un score de confiance de 0.9403.

```
<?xml version='1.0' encoding='utf-8'?>
<rdf:RDF xmlns='http://knowledgeweb.semanticweb.org/heterogeneity/alignment'
  xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
  xmlns:xsd='http://www.w3.org/2001/XMLSchema#'
  alignmentSource='AgreementMakerLight'>
  <Alignment>
    <xml>yes</xml>
    <level>0</level>
    <type>l1</type>
    <ontol>http://spider.sigappfr.org/HSSNdoc/hssn.ttl</ontol>
    <onto2>http://www.semanticweb.org/nabil/ontologies/2021/HR-Ontology</onto2>
    <uril>http://spider.sigappfr.org/HSSNdoc/hssn.ttl</uril>
    <uri2>http://www.semanticweb.org/nabil/ontologies/2021/HR-Ontology</uri2>
    <map>
      </map>
    </map>
    <map>
      <Cell>
        <entity1 rdf:resource="http://www.w3.org/2006/time#Instant"/>
        <entity2 rdf:resource="http://www.semanticweb.org/nabil/ontologies/2021/HR-Ontology#TimeInstant"/>
        <measure rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.9403</measure>
        <relation>=</relation>
      </Cell>
    </map>
  </map>
</rdf:RDF>
```

FIGURE A.1 – Capture d'écran d'un fichier d'alignement retourné par l'outil AgreementMakerLight

Les outils **AgreementMakerLight** et **TheAlignmentAPI** combinent tous les deux différentes techniques d'alignement [52, 37]. Ils sont développés en Java. Par conséquent, pour les intégrer dans la plateforme ISEEapp nous avons utilisé la bibliothèque Py4J⁴. La bibliothèque Py4J permet d'appeler les méthodes Java depuis un programme Python. Notez ici que le module d'alignement n'est pas accessible directement par l'utilisateur, il est utilisé implicitement par le module d'explication d'événements pour obtenir les résultats d'alignement.

Après avoir détaillé le module d'alignement des ontologies, la section suivante se consacre au module d'évaluation des alignements.

2. <https://github.com/AgreementMakerLight/AML-Project>
3. <https://moex.gitlabpages.inria.fr/alignapi/>
4. <https://github.com/AgreementMakerLight/AML-Project>

A.4 Évaluation des alignements

Ce module permet à l'utilisateur d'évaluer le résultat retourné par un outil d'alignement est de savoir si ce dernier lui permettra d'obtenir de bonnes explications. Pour ce faire, nous nous appuyons sur la métrique présentée dans la section 3.5.2.

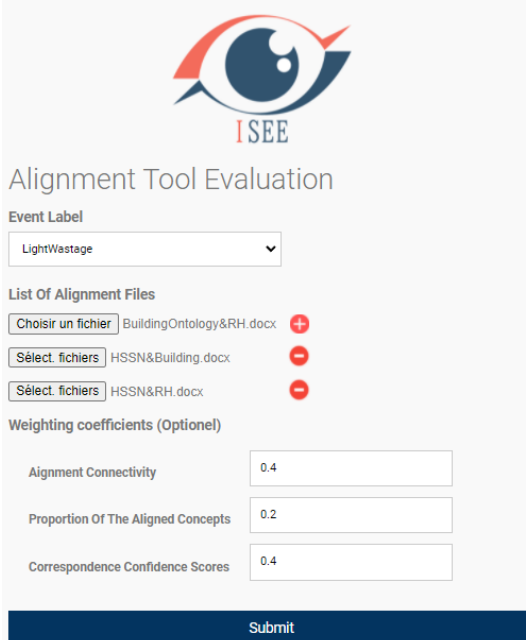


FIGURE A.2 – Évaluation d'un alignement d'ontologie sur la plateforme ISEEapp

La figure A.2 montre l'exemple de l'évaluation d'un alignement par rapport à l'événement gaspillage de lumière. L'utilisateur choisit l'événement qu'il souhaite expliquer (champ *Event Label*, figure A.2). Ensuite, il charge les différents fichiers d'alignement (champ *List Of Alignment Files*, figure A.2). Enfin, l'utilisateur précise les coefficients d'alignement qu'il veut accorder aux trois composantes de la métrique d'évaluation (la connectivité, la proportion des concepts alignés et le score de confiance des correspondances, *Weighting Coefficients* dans la figure A.2). Ensuite, le module calcule et retourne le score de l'alignement.

Pour développer la fonction du calcul de score nous nous sommes basés sur le paquet *Owlready2* précédemment cité pour manipuler les ontologies et sur le module *re*⁵ pour analyser les fichiers retournés par l'outil d'alignement. Le module *re* permet d'utiliser des expressions régulières avec Python.

Après avoir présenté le module d'alignement des ontologies, dans la section suivante nous passons au module de définition des événements.

5. <https://docs.python.org/3/library/re.html>

FIGURE A.3 – Choix des ontologies avant la définition d'un événement

FIGURE A.4 – Chargement d'une nouvelle ontologie par l'utilisateur dans la plateforme ISEEapp

A.5 Définition des événements

La plateforme ISEEapp propose à l'utilisateur de définir les événements qu'il souhaite détecter dans l'environnement à travers un formulaire. Ce dernier est basé sur la modélisation d'événements présentée dans le chapitre 3 (déf. 11, 12 et 13). L'utilisateur choisit tout d'abord les ontologies qu'il souhaite s'appuyer dessus pour définir l'événement à travers un premier formulaire (figure A.3). Celui-ci permet à l'utilisateur de choisir une ou plusieurs ontologies parmi celles enregistrées dans le système. Dans le cas où l'utilisateur souhaite employer une ontologie qui n'apparaît pas dans la liste, ce dernier a la possibilité de charger de nouvelles ontologies (la remarque au-dessous du bouton *Go to the next step*, figure A.3). L'utilisateur est alors redirigé vers la page présentée dans la figure A.4. Il doit tout d'abord saisir le label qu'il souhaite donner à l'ontologie, ensuite, l'utilisateur a le choix entre charger une ontologie depuis un fichier local (la case à gauche, figure A.4) ou bien à partir de l'URI de l'ontologie (la case à droite, figure A.4).

Après la validation de l'étape du choix des ontologies, l'utilisateur est redirigé vers le formulaire de définition d'évènement (figure A.5). Le formulaire est composé de deux parties principales, une première partie où l'utilisateur doit saisir le label de l'évènement (champ *Label*, figure A.5) et choisir le niveau d'urgence qu'il souhaite lui attribuer (champ *Emergency*, figure A.5). La deuxième partie du formulaire concerne l'espace de l'évènement (déf. 13). Pour chaque dimension *Feature*, *Sources*, *Time* et *Location*, l'utilisateur doit tout d'abord choisir une ou plusieurs ontologies de la liste des ontologies qu'il avait sélectionné dans l'étape précédente (champ *Ontology*, figure A.5). Le fait de sélectionner des ontologies permet à la plateforme de faire de l'auto-complétion lorsque l'utilisateur saisit les attributs de la dimension (Champs *Origin concept*, *Related Concepts* et *Related relations*, figure A.5). Après la validation de cette

étape, une fonction python se charge d’instancier l’ontologie HSSN-étendue avec la définition de l’événement.

The screenshot shows the ISEEAPP interface for defining an event. The form is organized as follows:

- Label:** Light Wastage
- Emergency:** Low
- eSpace:**
 - Features:**
 - Ontology: hssn
 - Origin concept: Luminosity
 - Related concepts: LightingSystem
 - Related relations: hasScreenLuminosity
 - Constraints: Luminosa > 50
 - Sources:**
 - Ontology: hssn
 - Origin concept: LightSensor
 - Related concepts: LightingSystem (optional)
 - Related relations: hasScreenLuminosity (optional)
 - Constraints: Click to add a constraint
 - Time:**
 - Ontology: hssn
 - Origin concept: TimeInstant
 - Related concepts: TemporalEntity
 - Related relations: hasTimeInstantInside
 - Constraints: TimeInsta inside [20:30:00]
 - Location:**
 - Ontology: hssn
 - Origin concept: Location
 - Related concepts: BuildingSpace
 - Related relations: aLocation
 - Constraints: Click to add a constraint

FIGURE A.5 – Définition d’un événement dans la plateforme ISEEapp

Après avoir présenté le module de définition des évènements, dans la section suivante nous passons au module de détection des événements.

A.6 Détection des événements

Ce module a pour objectif de détecter les occurrences des évènements définis par l’utilisateur. Pour ce faire, le module parcourt les données du réseau de capteurs (*données du réseau de capteurs*, figure 4.2) et vérifie les conditions de déclenchement précisées par l’utilisateur dans la définition de chaque événement (déf. 12). Les occurrences de déclenchement des événements sont alors ajoutées dans le graphe sémantique et plus précisément dans l’attribut I de l’espace de chaque événement (déf. 13). Bien entendu, ces occurrences de déclenchement d’évènements sont également affichées à l’utilisateur afin qu’il puisse demander leurs explications (section A.7). Pour développer la fonction de détection des événements nous nous sommes appuyés principalement sur le module *Owlready2* cité précédemment pour parcourir les définitions des événements et les données des capteurs dans le graphe sémantique. Pour comparer les données de capteurs avec les conditions de déclenchement, nous avons utilisé les opérateurs classiques de Python. Notez ici que si l’utilisateur définit un nouvel événement, la plateforme réexécute automatiquement le processus pour détecter les occurrences potentielles de cet événement. De même, si de nouvelles données de capteurs intègrent le système le processus de détection des événements est réexécuté sur ces données.

Après avoir présenté le module de détection des événements, dans la section suivante nous passons au module d'explication d'événements.

A.7 Explication d'événements

Pour obtenir une explication, l'utilisateur doit d'abord sélectionner l'événement qu'il souhaite expliquer parmi ceux détectés par le système et l'outil à utiliser pour aligner les ontologies de domaine. La plateforme ISEEapp utilise par défaut l'outil d'alignement *AgreementMakerLight* [52] étant donné qu'il nous a permis d'obtenir les meilleurs résultats lors de la phase d'évaluation (chapitre 4.1). Néanmoins, l'utilisateur peut choisir d'employer un autre outil parmi ceux intégrés dans le système. Pour aider l'utilisateur à choisir l'outil d'alignement qui lui convient le mieux, la plateforme calcule automatiquement le score de l'outil en faisant appel au module d'évaluation des alignements (section A.3), puis l'affiche à l'utilisateur.

Une fois le choix de l'utilisateur validé, le module d'explication des événements peut commencer à construire l'explication. Ce dernier exécute puis retourne l'explication à l'utilisateur. La fonction de recherche d'explication mis en oeuvre les trois étapes du processus ISEE (étapes (1), (2) et (3), figure 3.1). La figure A.6 montre un exemple d'explication complète retournée pour une occurrence de l'événement gaspillage de lumière. Les attributs *What*, *Who*, *When*, *Where* *How* détaillent respectivement l'identifiant de l'événement déclenché, le capteur qui a détecté l'événement, l'heure du déclenchement, le lieu du déclenchement et les observations qui ont permis de détecter l'événement. Enfin l'attribut *Why_i* contient plusieurs tuples *Why_i* (nous n'affichons pas ici les indices pour des raisons de clarté). Le tuple visible sur la capture suggère que l'employé Roland Perrin a peut-être laissé les lampes 457 et 458 allumées dans le bureau 400 quand il a quitté le bureau à 18:20:09. Le score de pertinence de ce tuple *Why_i* est de 0.813 (*Relevance score*, figure A.6). Nous rappelons ici que cette explication est structurée selon notre modèle détaillé dans le chapitre 3 (section 16). Le score de pertinence est calculé en utilisant la métrique proposée dans la section 3.5.3.3.

Pour développer la fonction de construction des explications, nous nous sommes appuyés sur le paquet *Owlready2* et sur le module *re* cités plus haut respectivement pour manipuler les ontologies et pour analyser les fichiers retournés par le module d'alignement des ontologies (section A.3). Pour comparer les instances spatiales et temporelles nous avons utilisé respectivement les deux modules *difflib*⁶ et *time*⁷.

A.8 Conclusion

Dans cet annexe nous avons présenté la plateforme ISEEapp pour l'explication des événements dans les environnements hybrides. Cette plateforme mis en oeuvre notre

6. <https://docs.python.org/3/library/difflib.html>

7. <https://docs.python.org/fr/3/library/time.html>

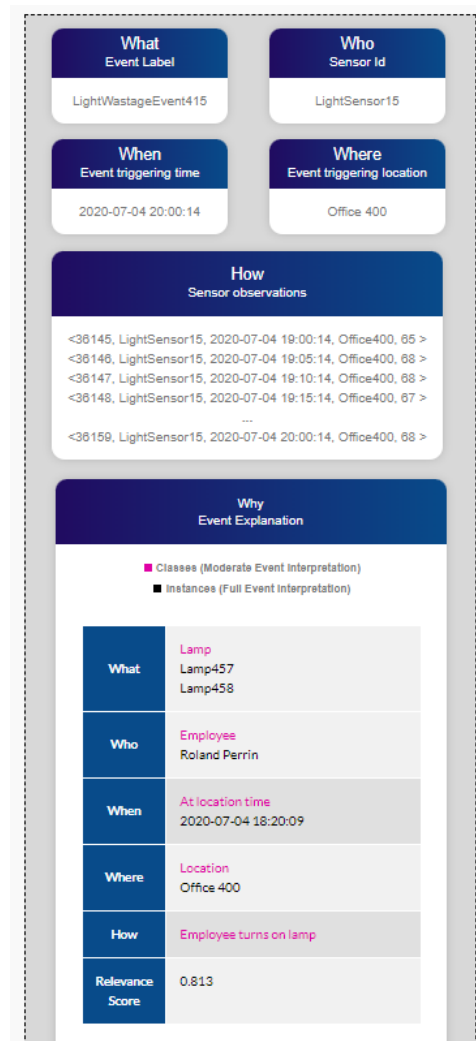


FIGURE A.6 – Exemple d’explication d’un évènement gaspillage de lumière dans la plateforme ISEEapp

proposition, le système ISEE qui a été détaillé dans le chapitre 3. La plateforme ISEEapp est développée en langage Python en utilisant le micro cadre (framework) open-source Flask⁸. L’interface graphique de la plateforme est développée en HTML, CSS et JavaScript. Le dépôt de données de la plateforme est composé de cinq jeux de données : données du réseau de capteurs, corpus documentaire, définitions des composantes de l’environnement, explications des événements passés et ontologies de domaine. La plateforme ISEEapp est constituée de deux couches, une couche services et une couche accès aux ressources. La couche services comprend six modules organisés en trois catégories : (i) intégration des connaissances de domaine, (ii) gestion des événements et (iii) processus ISEE. La catégorie d’intégration des connaissances de domaine

8. <https://flask.palletsprojects.com/en/2.1.x/>

comprend trois modules qui correspondent aux étapes du processus d'intégration des connaissances de domaine, à savoir **le module d'instanciation des ontologies de domaine** qui s'appuie sur les différents jeux de données pour instancier les ontologies de domaine; **Le module d'alignement des ontologies de domaine** qui intègre deux outils d'alignement (AgreementMakerLight [52] et TheAlignmentAPI [37]); **Le module d'évaluation des alignements** qui renvoie à l'utilisateur un score indiquant si l'alignement établi lui permettra éventuellement d'obtenir de bonnes explications. Ensuite, la catégorie de gestion des événements comprend deux modules, à savoir **le module de la définition des événements** qui permet à l'utilisateur de définir les événements qu'il souhaite détecter selon le modèle proposé dans le chapitre précédent (la section 3.4.1); **Le module de détection des événements** qui analyse les données de capteurs et détecte les occurrences des événements définis par l'utilisateur. Enfin, la catégorie processus ISEE comporte un seul module qui englobe les différentes étapes du processus ISEE. Ce module permet à l'utilisateur de choisir l'événement qu'il souhaite expliquer, puis cherche l'explication et la retourne à l'utilisateur. La couche d'accès aux ressources comprend les fonctions python pour la lecture et l'écriture dans le dépôt de données.

Pour développer les différents modules de la couche services, nous nous sommes appuyés principalement sur le paquet json⁹ pour manipuler les fichiers JSON, le paquet Owlready2 pour charger et manipuler des ontologies et le module re¹⁰ pour analyser les fichiers retournés par l'outil d'alignement.

9. <https://docs.python.org/3/library/json.html>

10. <https://docs.python.org/3/library/re.html>

Bibliographie

- [1] World Wide Web Consortium (W3C). resource description framework. <https://www.w3.org/RDF/>. Accessed : 2022-04-01.
- [2] ABUBAKAR, M., HAMDAN, H., MUSTAPHA, N., AND ARIS, T. N. M. Instance-based ontology matching : A literature review. In *Recent Advances on Soft Computing and Data Mining - Proceedings of the Third International Conference on Soft Computing and Data Mining (SCDM 2018), Johor, Malaysia, February 06-07, 2018* (2018), R. Ghazali, M. M. Deris, N. M. Nawawi, and J. H. Abawajy, Eds., vol. 700 of *Advances in Intelligent Systems and Computing*, Springer, pp. 455–469.
- [3] AHMED, E., YAQOOB, I., GANI, A., RAZZAK, M. I., AND GUIZANI, M. Internet-of-things-based smart environments : state of the art, taxonomy, and open research challenges. *IEEE Wirel. Commun.* 23, 5 (2016), 10–16.
- [4] ALAM, M. R., REAZ, M. B. I., AND ALI, M. A. M. A review of smart homes - past, present, and future. *IEEE Trans. Syst. Man Cybern. Part C* 42, 6 (2012), 1190–1203.
- [5] ALLAM, A. M. N., AND HAGGAG, M. H. The question answering systems : A survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS)* 2, 3 (2012).
- [6] ALTENBERG, B. Causal linking in spoken and written english. *Studia linguistica* 38, 1 (1984), 20–69.
- [7] AMIRIAN, S., AND MOHAMMADI, S. Ontology alignment using wordnet method. *Int. J. Comput. Sci. Netw. Secur.* 17, 7 (2017), 161–167.
- [8] ANGSUCHOTMETEE, C., CHBEIR, R., AND CARDINALE, Y. Mssn-onto : An ontology-based approach for flexible event processing in multimedia sensor networks. *Future Gener. Comput. Syst.* 108 (2020), 1140–1158.
- [9] ARASU, A., BABU, S., AND WIDOM, J. CQL : A language for continuous queries over streams and relations. In *Database Programming Languages, 9th International Workshop, DBPL 2003, Potsdam, Germany, September 6-8, 2003, Revised Papers* (2003), G. Lausen and D. Suciu, Eds., vol. 2921 of *Lecture Notes in Computer Science*, Springer, pp. 1–19.
- [10] ARDJANI, F., BOUCHIHA, D., AND MALKI, M. Ontology-alignment techniques : Survey and analysis. *International Journal of Modern Education & Computer Science* 7, 11 (2015).

- [11] ARORA, M., KANJILAL, U., AND VARSHNEY, D. Efficient and intelligent information retrieval using support vector machine (svm). *Int. J. Soft Comput. Eng. (IJSCE)* 1, 6 (2012), 39–43.
- [12] ATENCIA, M., DAVID, J., AND EUZENAT, J. Data interlinking through robust linkkey extraction. In *ECAI 2014 - 21st European Conference on Artificial Intelligence, 18-22 August 2014, Prague, Czech Republic - Including Prestigious Applications of Intelligent Systems (PAIS 2014)* (2014), T. Schaub, G. Friedrich, and B. O’Sullivan, Eds., vol. 263 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, pp. 15–20.
- [13] ATENCIA, M., DAVID, J., EUZENAT, J., NAPOLI, A., AND VIZZINI, J. Link key candidate extraction with relational concept analysis. *Discret. Appl. Math.* 273 (2020), 2–20.
- [14] AUNIMO, L., KUUSKOSKI, R., AND MAKKONEN, J. Finnish as source language in bilingual question answering. In *Workshop of the Cross-Language Evaluation Forum for European Languages* (2004), Springer, pp. 482–493.
- [15] AZAD, H. K., AND DEEPAK, A. A new approach for query expansion using wikipedia and wordnet. *Inf. Sci.* 492 (2019), 147–163.
- [16] BADER, A., KOPP, O., AND FALKENTHAL, M. Survey and comparison of open source time series databases. In *Datenbanksysteme für Business, Technologie und Web (BTW 2017), 17. Fachtagung des GI-Fachbereichs „Datenbanken und Informationssysteme“ (DBIS), 6.-10. März 2017, Stuttgart, Germany, Workshopband* (2017), B. Mitschang, N. Ritter, H. Schwarz, M. Klettke, A. Thor, O. Kopp, and M. Wieland, Eds., vol. P-266 of *LNI, GI*, pp. 249–268.
- [17] BAEZA-YATES, R., RIBEIRO-NETO, B., ET AL. *Modern information retrieval*, vol. 463. ACM press New York, 1999.
- [18] BAJAJ, G., AGARWAL, R., SINGH, P., GEORGANTAS, N., AND ISSARNY, V. A study of existing ontologies in the iot-domain. *CoRR abs/1707.00112* (2017).
- [19] BANERJEE, S., AND MISHRA, A. Semantic exploration of sensor data. In *Proceedings of the 5th International Workshop on Web-scale Knowledge Representation Retrieval & Reasoning, Web-KR@CIKM 2014, Shanghai, China, November 3, 2014* (2014), Y. Zeng, S. Kotoulas, and Z. Huang, Eds., ACM, pp. 55–58.
- [20] BAQA, H., BAUER, M., BILBAO, S., CORCHERO, A., DANIELE, L., ESNAOLA, I., FERNÁNDEZ, I., FRÄNBERG, Ö., GARCÍA CASTRO, R., GIROD-GENET, M., GUILLEMIN, P., GYRARD, A., EL KAED, C., KUNG, A., LEE, J., LEFRANÇOIS, M., LI, W., RAGGETT, D., AND WETTERWALD, M. *Semantic IoT Solutions -A Developer Perspective*. Oct. 2019.
- [21] BOUADJENEK, M. R. *Infrastructure and Algorithms for Information Retrieval Based On Social Network Analysis/Mining. (Infrastructure et Algorithmes pour la Recherche d’Information Basés sur l’Analyse des Réseaux Sociaux)*. PhD thesis, University of Paris-Saclay, France, 2013.

- [22] BUSCALDI, D., ROSSO, P., SORIANO, J. M. G., AND SANCHIS, E. Answering questions with an n -gram based passage retrieval engine. *J. Intell. Inf. Syst.* 34, 2 (2010), 113–134.
- [23] BÜTTCHER, S., CLARKE, C. L., AND CORMACK, G. V. *Information retrieval : Implementing and evaluating search engines*. Mit Press, 2016.
- [24] CALBIMONTE, J., JEUNG, H., CORCHO, Ó., AND ABERER, K. Enabling query technologies for the semantic sensor web. *Int. J. Semantic Web Inf. Syst.* 8, 1 (2012), 43–63.
- [25] CAO, Y., XU, J., LIU, T.-Y., LI, H., HUANG, Y., AND HON, H.-W. Adapting ranking svm to document retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (2006), pp. 186–193.
- [26] CATARCI, T., DONGILLI, P., MASCIO, T. D., FRANCONI, E., SANTUCCI, G., AND TESSARIS, S. An ontology based visual tool for query formulation support. In *Proceedings of the 16th European Conference on Artificial Intelligence, ECAI'2004, including Prestigious Applicants of Intelligent Systems, PAIS 2004, Valencia, Spain, August 22-27, 2004* (2004), R. L. de Mántaras and L. Saitta, Eds., IOS Press, pp. 308–312.
- [27] CERÓN-FIGUEROA, S., LÓPEZ-YÁÑEZ, I., ALHALABI, W., NIETO, O. C., VILLUENDAS-REY, Y., ALDAPE-PÉREZ, M., AND YÁÑEZ-MÁRQUEZ, C. Instance-based ontology matching for e-learning material using an associative pattern classifier. *Comput. Hum. Behav.* 69 (2017), 218–225.
- [28] CHAWLA, S. Semantic query expansion using cluster based domain ontologies. *Int. J. Inf. Retr. Res.* 2, 2 (2012), 13–28.
- [29] CHEATHAM, M., AND HITZLER, P. String similarity metrics for ontology alignment. In *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II* (2013), H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. F. Noy, C. Welty, and K. Janowicz, Eds., vol. 8219 of *Lecture Notes in Computer Science*, Springer, pp. 294–309.
- [30] CHENIKI, N., BELKHIR, A., SAM, Y., AND MESSAI, N. LODS : A linked open data based similarity measure. In *25th IEEE International Conference on Enabling Technologies : Infrastructure for Collaborative Enterprises, WETICE 2016, Paris, France, June 13-15, 2016* (2016), S. Reddy and W. Gaaloul, Eds., IEEE Computer Society, pp. 229–234.
- [31] COMPTON, M., BARNAGHI, P. M., BERMUDEZ, L., GARCIA-CASTRO, R., CORCHO, Ó., COX, S. J. D., GRAYBEAL, J., HAUSWIRTH, M., HENSON, C. A., HERZOG, A., HUANG, V. A., JANOWICZ, K., KELSEY, W. D., PHUOC, D. L., LEFORT, L., LEGGIERI, M., NEUHAUS, H., NIKOLOV, A., PAGE, K. R., PASSANT, A., SHETH, A. P., AND TAYLOR, K. The SSN ontology of the W3C semantic sensor network incubator group. *J. Web Semant.* 17 (2012), 25–32.

- [32] COMPTON, M., HENSON, C. A., NEUHAUS, H., LEFORT, L., AND SHETH, A. P. A survey of the semantic specification of sensors. In *Proceedings of the 2nd International Workshop on Semantic Sensor Networks (SSN09), collocated with the 8th International Semantic Web Conference (ISWC-2009), Washington DC, USA, October 26, 2009* (2009), K. Taylor and D. D. Roure, Eds., vol. 522 of *CEUR Workshop Proceedings*, CEUR-WS.org, pp. 17–32.
- [33] COOK, D. J., AND DAS, S. K. *Smart environments - technology, protocols and applications*. Wiley, 2005.
- [34] CROFT, W. B., AND THOMPSON, R. H. I3r : A new approach to the design of document retrieval systems. *J. Am. Soc. Inf. Sci.* 38, 6 (nov 1987), 389–404.
- [35] CUNNINGHAM, H. Information extraction, automatic. *Encyclopedia of language and linguistics*, 3, 8 (2005), 10.
- [36] DAMLJANOVIC, D., AGATONOVIC, M., AND CUNNINGHAM, H. Identification of the question focus : Combining syntactic analysis and ontology-based lookup through the user interaction. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta* (2010), N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds., European Language Resources Association.
- [37] DAVID, J., EUZENAT, J., SCHARFFE, F., AND DOS SANTOS, C. T. The alignment API 4.0. *Semantic Web 2*, 1 (2011), 3–10.
- [38] DE FARIA, C. G., SERRA, I., AND GIRARDI, R. A domain-independent process for automatic ontology population from text. *Sci. Comput. Program.* 95 (2014), 26–43.
- [39] DE PAOLIS, L. T., DE LUCA, V., AND PAIANO, R. Sensor data collection and analytics with thingsboard and spark streaming. In *2018 IEEE workshop on environmental, energy, and structural monitoring systems (EESMS)* (2018), IEEE, pp. 1–6.
- [40] DEEPAK, G., AND PRIYADARSHINI, S. Personalized and enhanced hybridized semantic algorithm for web image retrieval incorporating ontology classification, strategic query expansion, and content-based analysis. *Comput. Electr. Eng.* 72 (2018), 14–25.
- [41] DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K., AND HARSHMAN, R. Indexing by latent semantic analysis. *Journal of the American society for information science* 41, 6 (1990), 391–407.
- [42] DIALLO, G. An effective method of large scale ontology matching. *J. Biomed. Semant.* 5 (2014), 44.
- [43] DIMITRAKIS, E., SGONTZOS, K., AND TZITZIKAS, Y. A survey on question answering systems over linked data and documents. *J. Intell. Inf. Syst.* 55, 2 (2020), 233–259.
- [44] DIMOU, A., SANDE, M. V., COLPAERT, P., VERBORGH, R., MANNENS, E., AND DE WALLE, R. V. RML : A generic language for integrated RDF mappings of heterogeneous data. In *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014)*,

- Seoul, Korea, April 8, 2014* (2014), C. Bizer, T. Heath, S. Auer, and T. Berners-Lee, Eds., vol. 1184 of *CEUR Workshop Proceedings*, CEUR-WS.org.
- [45] DJEDDI, W. E., AND KHADIR, T. A dynamic multistrategy ontology alignment framework based on semantic relationship using wordnet. In *Proceedings of the Third International Conference on Computer Science and its Applications (CIIA'11), Saida, Algeria, December 13-15, 2011* (2011), A. Amine, O. A. Mohamed, B. Benatallah, and Z. Elberrichi, Eds., vol. 825 of *CEUR Workshop Proceedings*, CEUR-WS.org.
- [46] E COMMERCE, U. A. Digitalisation des entreprises : Le virage numérique, c'est maintenant!, 2021.
- [47] ECHIHABI, A., HERMJAKOB, U., HOVY, E., MARCU, D., MELZ, E., AND RAVICHANDRAN, D. How to select an answer string? In *Advances in open domain question answering*. Springer, 2008, pp. 383–406.
- [48] ESPOSITO, M., DAMIANO, E., MINUTOLO, A., PIETRO, G. D., AND FUJITA, H. Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering. *Inf. Sci.* 514 (2020), 88–105.
- [49] EUZENAT, J., AND SHVAIKO, P. *Ontology matching*. Springer, 2007.
- [50] FAN, Z. *Concise Pattern Learning for RDF Data Sets Interlinking. (Apprentissage de Motifs Concis pour le Liage de Données RDF)*. PhD thesis, University of Grenoble, France, 2014.
- [51] FARHOODI, M., MAHMOUDI, M., BIDOKI, A. Z., YARI, A., AND AZADNIA, M. Query expansion using persian ontology derived from wikipedia. *World Applied Sciences Journal* 7, 4 (2009), 410–417.
- [52] FARIA, D., PESQUITA, C., SANTOS, E., PALMONARI, M., CRUZ, I. F., AND COUTO, F. M. The agreementmakerlight ontology matching system. In *On the Move to Meaningful Internet Systems : OTM 2013 Conferences - Confederated International Conferences : CoopIS, DOA-Trusted Cloud, and ODBASE 2013, Graz, Austria, September 9-13, 2013. Proceedings* (2013), R. Meersman, H. Panetto, T. S. Dillon, J. Eder, Z. Bellahsene, N. Ritter, P. D. Leenheer, and D. Dou, Eds., vol. 8185 of *Lecture Notes in Computer Science*, Springer, pp. 527–541.
- [53] FATIMA, H., AND WASNIK, K. Comparison of sql, nosql and newsql databases for internet of things. In *2016 IEEE Bombay Section Symposium (IBSS)* (2016), IEEE, pp. 1–6.
- [54] FERNÁNDEZ, M., CANTADOR, I., LÓPEZ, V., VALLET, D., CASTELLS, P., AND MOTTA, E. Semantically enhanced information retrieval : An ontology-based approach. *J. Web Semant.* 9, 4 (2011), 434–452.
- [55] FERNÁNDEZ, M., CANTADOR, I., LÓPEZ, V., VALLET, D., CASTELLS, P., AND MOTTA, E. Semantically enhanced information retrieval : An ontology-based approach. *Journal of Web Semantics* 9, 4 (2011), 434–452.

- [56] FERNANDEZ, R. C., AND MADDEN, S. Termite : a system for tunneling through heterogeneous data. In *Proceedings of the Second International Workshop on Exploiting Artificial Intelligence Techniques for Data Management, aiDM@SIGMOD 2019, Amsterdam, The Netherlands, July 5, 2019* (2019), R. Bordawekar and O. Shmueli, Eds., ACM, pp. 7 :1–7 :8.
- [57] FERNÁNDEZ, S., AND ITO, T. Semantic integration of sensor data with SSN ontology in a multi-agent architecture for intelligent transportation systems. *IEICE Trans. Inf. Syst.* 100-D, 12 (2017), 2915–2922.
- [58] FOTSOH, A. *Recherche d'entités nommées complexes sur le Web - propositions pour l'extraction et pour le calcul de similarité*. Theses, Université de Pau et des Pays de l'Adour, Feb. 2018.
- [59] FOX, E. A., AND FRANCE, R. K. Architecture of an expert system for composite document analysis, representation, and retrieval. *Int. J. Approx. Reason.* 1 (1987), 151–175.
- [60] GALIĆ, Z., MEŠKOVIĆ, E., KRIŽANOVIĆ, K., AND BARANOVIĆ, M. Oceanus : a spatio-temporal data stream system prototype. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on GeoStreaming* (2012), pp. 109–115.
- [61] GARCÍA, E., AND SICILIA, M.-Á. Designing ontology-based interactive information retrieval interfaces. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (2003), Springer, pp. 152–165.
- [62] GARCÍA-GONZÁLEZ, H., FERNÁNDEZ-ÁLVAREZ, D., AND GAYO, J. E. L. Shexml : An heterogeneous data mapping language based on shex. In *Proceedings of the EKAW 2018 Posters and Demonstrations Session co-located with 21st International Conference on Knowledge Engineering and Knowledge Management (EKAW 2018), Nancy, France, November 12-16, 2018* (2018), P. Cimiano and O. Corby, Eds., vol. 2262 of *CEUR Workshop Proceedings*, CEUR-WS.org, pp. 9–12.
- [63] GIUNCHIGLIA, F., YATSKEVICH, M., AND GIUNCHIGLIA, E. Efficient semantic matching. In *The Semantic Web : Research and Applications, Second European Semantic Web Conference, ESWC 2005, Heraklion, Crete, Greece, May 29 - June 1, 2005, Proceedings* (2005), A. Gómez-Pérez and J. Euzenat, Eds., vol. 3532 of *Lecture Notes in Computer Science*, Springer, pp. 272–289.
- [64] GLÜCKSTAD, F. K. Terminological ontology and cognitive processes in translation. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, PACLIC 24, Tohoku University, Japan, 4-7 November 2010* (2010), R. Ootoguro, K. Ishikawa, H. Umemoto, K. Yoshimoto, and Y. Harada, Eds., Institute for Digital Enhancement of Cognitive Development, Waseda University, pp. 629–636.
- [65] GOKER, A., AND DAVIES, J. *Information retrieval : Searching in the 21st century*. John Wiley & Sons, 2009.
- [66] GONG, Z., CHEANG, C. W., AND U, L. H. Multi-term web query expansion using wordnet. In *Database and Expert Systems Applications, 17th International*

- Conference, DEXA 2006, Kraków, Poland, September 4-8, 2006, Proceedings* (2006), S. Bressan, J. Küng, and R. R. Wagner, Eds., vol. 4080 of *Lecture Notes in Computer Science*, Springer, pp. 379–388.
- [67] GREENE, D., AND CUNNINGHAM, P. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006* (2006), W. W. Cohen and A. W. Moore, Eds., vol. 148 of *ACM International Conference Proceeding Series*, ACM, pp. 377–384.
- [68] GREENGARD, S. *The internet of things*. MIT press, 2021.
- [69] GUO, K., LIANG, Z., TANG, Y., AND CHI, T. SOR : an optimized semantic ontology retrieval algorithm for heterogeneous multimedia big data. *J. Comput. Sci.* 28 (2018), 455–465.
- [70] HAHM, G. J., YI, M. Y., LEE, J., AND SUH, H. A personalized query expansion approach for engineering document retrieval. *Adv. Eng. Informatics* 28, 4 (2014), 344–359.
- [71] HALLER, A., JANOWICZ, K., COX, S. J. D., LEFRANÇOIS, M., TAYLOR, K., PHUOC, D. L., LIEBERMAN, J., GARCÍA-CASTRO, R., ATKINSON, R., AND STADLER, C. The modular SSN ontology : A joint W3C and OGC standard specifying the semantics of sensors, observations, sampling, and actuation. *Semantic Web* 10, 1 (2019), 9–32.
- [72] HAMBORG, F., BREITINGER, C., AND GIPP, B. Giveme5w1h : A universal system for extracting main events from news articles. In *Proceedings of the 7th International Workshop on News Recommendation and Analytics in conjunction with 13th ACM Conference on Recommender Systems, INRA@RecSys 2019, Copenhagen, Denmark, September 20, 2019* (2019), Ö. Özgöbek, B. Kille, J. A. Gulla, and A. Lommatzsch, Eds., vol. 2554 of *CEUR Workshop Proceedings*, CEUR-WS.org.
- [73] HAMBORG, F., LACHNIT, S., SCHUBOTZ, M., HEPP, T., AND GIPP, B. Giveme5w : Main event retrieval from news articles by extraction of the five journalistic W questions. In *Transforming Digital Worlds - 13th International Conference, iConference 2018, Sheffield, UK, March 25-28, 2018, Proceedings* (2018), G. Chowdhury, J. McLeod, V. J. Gillet, and P. Willett, Eds., vol. 10766 of *Lecture Notes in Computer Science*, Springer, pp. 356–366.
- [74] HAMBORG, F., MEUSCHKE, N., AND GIPP, B. Matrix-based news aggregation : Exploring different news perspectives. In *2017 ACM/IEEE Joint Conference on Digital Libraries, JCDL 2017, Toronto, ON, Canada, June 19-23, 2017* (2017), IEEE Computer Society, pp. 69–78.
- [75] HAMED, S. K., AND AZIZ, M. J. A. A question answering system on holy quran translation based on question expansion technique and neural network classification. *J. Comput. Sci.* 12, 3 (2016), 169–177.
- [76] HENDRICKSON, B. Latent semantic analysis and fiedler retrieval. *Linear Algebra and its Applications* 421, 2-3 (2007), 345–355.

- [77] HOOI, Y. K., HASSAN, M. F., AND SHARIFF, A. M. A survey on ontology mapping techniques. *Advances in Computer Science and its Applications* (2014), 829–836.
- [78] HU, Y., WU, Z., AND GUO, M. Ontology driven adaptive data processing in wireless sensor networks. In *Proceedings of the 2nd International Conference on Scalable Information Systems, Infoscale 2007, Suzhou, China, June 6-8, 2007* (2007), J. Li, W. Lee, and F. Silvestri, Eds., vol. 304 of *ACM International Conference Proceeding Series*, ACM, p. 46.
- [79] HUANG, B., BOUGUETTAYA, A., AND NEIAT, A. G. Discovering spatio-temporal relationships among iot services. In *2018 IEEE International Conference on Web Services (ICWS)* (2018), IEEE, pp. 347–350.
- [80] IRFAN, S., AND BABU, B. Information retrieval in big data using evolutionary computation : A survey. In *2016 International conference on computing, communication and automation (ICCCA)* (2016), IEEE, pp. 208–213.
- [81] ISELE, R., AND BIZER, C. Active learning of expressive linkage rules using genetic programming. *J. Web Semant.* 23 (2013), 2–15.
- [82] JAUREGUIBERRY, F. De l’usage des technologies de l’information et de la communication comme apprentissage créatif. *Education et sociétés* 22 (01 2009).
- [83] JENSEN, S. K., PEDERSEN, T. B., AND THOMSEN, C. Time series management systems : A survey. *IEEE Transactions on Knowledge and Data Engineering* 29, 11 (2017), 2581–2600.
- [84] JEONG, S., ZHANG, Y., O’CONNOR, S., LYNCH, J. P., SOHN, H., AND LAW, K. H. A nosql data management infrastructure for bridge monitoring. *Smart Structures and Systems* 17, 4 (2016), 669–690.
- [85] JESSUP, E. R., AND MARTIN, J. H. Taking a new look at the latent semantic analysis approach to information retrieval. *Computational information retrieval 2001* (2001), 121–144.
- [86] JIANG, S., LOWD, D., KAFLE, S., AND DOU, D. Ontology matching with knowledge rules. *Trans. Large Scale Data Knowl. Centered Syst.* 28 (2016), 75–95.
- [87] JIANG, X., HADID, A., PANG, Y., GRANGER, E., AND FENG, X. *Deep Learning in object detection and recognition*. Springer, 2019.
- [88] JIANG, Y. Semantically-enhanced information retrieval using multiple knowledge sources. *Cluster Computing* (2020), 1–20.
- [89] JIMÉNEZ-RUIZ, E., GRAU, B. C., AND ZHOU, Y. Logmap 2.0 : towards logic-based, scalable and interactive ontology matching. In *Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences, SWAT4LS 2011, London, United Kingdom, December 07-09, 2011* (2011), A. Paschke, A. Burger, P. Romano, M. S. Marshall, and A. Splendiani, Eds., ACM, pp. 45–46.
- [90] JIN-DE, T. Streamsql : A query language for stream data. *Computer Systems & Applications* (2010).

- [91] JOSLYN, C. A., PAULSON, P. R., AND WHITE, A. M. Measuring the structural preservation of semantic hierarchy alignment. In *Proceedings of the 4th International Workshop on Ontology Matching (OM-2009) collocated with the 8th International Semantic Web Conference (ISWC-2009) Chantilly, USA, October 25, 2009* (2009), P. Shvaiko, J. Euzenat, F. Giunchiglia, H. Stuckenschmidt, N. F. Noy, and A. Rosenthal, Eds., vol. 551 of *CEUR Workshop Proceedings*, CEUR-WS.org.
- [92] KAED, C. E., DANILCHENKO, V., DELPECH, F., BRODEUR, J., AND RADISSON, A. Linking an asset and a domain specific ontology for a simple asset timeseries application. In *IEEE International Conference on Big Data (IEEE BigData 2018), Seattle, WA, USA, December 10-13, 2018* (2018), N. Abe, H. Liu, C. Pu, X. Hu, N. K. Ahmed, M. Qiao, Y. Song, D. Kossmann, B. Liu, K. Lee, J. Tang, J. He, and J. S. Saltz, Eds., IEEE, pp. 4182–4188.
- [93] KEJRIWAL, M., AND MIRANKER, D. P. An unsupervised instance matcher for schema-free RDF data. *J. Web Semant.* 35 (2015), 102–123.
- [94] KHALIFI, H., ELQADI, A., AND GHANOU, Y. Support vector machines for a new hybrid information retrieval system. *Procedia Computer Science* 127 (2018), 139–145.
- [95] KHIAT, A., AND BENAÏSSA, M. AOT / AOTL results for OAEI 2014. In *Proceedings of the 9th International Workshop on Ontology Matching collocated with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Trentino, Italy, October 20, 2014* (2014), P. Shvaiko, J. Euzenat, M. Mao, E. Jiménez-Ruiz, J. Li, and A. Ngonga, Eds., vol. 1317 of *CEUR Workshop Proceedings*, CEUR-WS.org, pp. 113–119.
- [96] KHIAT, A., AND BENAÏSSA, M. Insmt / insmtl results for OAEI 2014 instance matching. In *Proceedings of the 9th International Workshop on Ontology Matching collocated with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Trentino, Italy, October 20, 2014* (2014), P. Shvaiko, J. Euzenat, M. Mao, E. Jiménez-Ruiz, J. Li, and A. Ngonga, Eds., vol. 1317 of *CEUR Workshop Proceedings*, CEUR-WS.org, pp. 120–125.
- [97] KHIAT, A., AND BENAÏSSA, M. A new instance-based approach for ontology alignment. *Int. J. Semantic Web Inf. Syst.* 11, 3 (2015), 25–43.
- [98] KHOO, C. S., KORNFILT, J., ODDY, R. N., AND MYAENG, S. H. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing* 13, 4 (1998), 177–186.
- [99] KIM, J., KWON, H., KIM, D., KWAK, H., AND LEE, S. J. Building a service-oriented ontology for wireless sensor networks. In *7th IEEE/ACIS International Conference on Computer and Information Science, IEEE/ACIS ICIS 2008, 14-16 May 2008, Portland, Oregon, USA* (2008), R. Y. Lee, Ed., IEEE Computer Society, pp. 649–654.
- [100] KIM, J.-J. Spatio-temporal sensor data processing techniques. *Journal of Information Processing Systems* 13, 5 (2017), 1259–1276.

- [101] KO, J., NYBERG, E., AND SI, L. A probabilistic graphical model for joint answer ranking in question answering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (2007), pp. 343–350.
- [102] KRIŽANOVIĆ, K., GALIĆ, Z., AND BARANOVIĆ, M. Spatio-temporal data streams : an approach to managing moving objects. In *The 33rd International Convention MIPRO* (2010), IEEE, pp. 744–749.
- [103] KWOK, K. L. A neural network for probabilistic information retrieval. In *SIGIR'89, 12th International Conference on Research and Development in Information Retrieval, Cambridge, Massachusetts, USA, June 25-28, 1989, Proceedings* (1989), N. J. Belkin and C. J. van Rijsbergen, Eds., ACM, pp. 21–30.
- [104] KWON, S., PARK, D., BANG, H., AND PARK, Y. Real-time and parallel semantic translation technique for large-scale streaming sensor data in an iot environment. *Journal of KIISE* 42, 1 (2015), 54–67.
- [105] LANDAUER, T. K., FOLTZ, P. W., AND LAHAM, D. An introduction to latent semantic analysis. *Discourse processes* 25, 2-3 (1998), 259–284.
- [106] LEE, J., YUN, S., KIM, H., KO, M., AND KANG, J. Ranking paragraphs for improving answer recall in open-domain question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018* (2018), E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds., Association for Computational Linguistics, pp. 565–569.
- [107] LEGRAIN, M. -A. La digitalisation ; ses avantages et ses outils, 2018.
- [108] LESKOVEC, J., GROBELNIK, M., AND MILIC-FRAYLING, N. Learning sub-structures of document semantic graphs for document summarization. In *LinkKDD Workshop* (2004), vol. 133, p. 138.
- [109] LONI, B. A survey of state-of-the-art methods on question classification.
- [110] LUBANI, M., NOAH, S. A. M., AND MAHMUD, R. Ontology population : Approaches and design aspects. *J. Inf. Sci.* 45, 4 (2019).
- [111] LUSSET, M. La digitalisation de l'entreprise, 2020.
- [112] LUUKKONEN, I., TOIVANEN, M., MURSU, A., SARANTO, K., AND KORPELA, M. Researching an activity-driven approach to information systems development. In *Handbook of Research on ICTs and Management Systems for Improving Efficiency in Healthcare and Social Care*. IGI Global, 2013, pp. 431–450.
- [113] MAKKI, J. Ontoprime : A prototype for automating ontology population. *International Journal of Web/Semantic Technology (IJWesT)* 8 (2017).
- [114] MANNING, C. D., RAGHAVAN, P., AND SCHÜTZE, H. *Boolean retrieval*. Cambridge University Press, 2008, p. 1–17.
- [115] MANSOUR, E. *Event detection in connected environments*. Theses, Université de Pau et des Pays de l'Adour, Nov. 2019.

- [116] MANSOUR, E., CHBEIR, R., AND ARNOULD, P. HSSN : an ontology for hybrid semantic sensor networks. In *Proceedings of the 23rd International Database Applications & Engineering Symposium, IDEAS 2019, Athens, Greece, June 10-12, 2019* (2019), B. C. Desai, D. Anagnostopoulos, Y. Manolopoulos, and M. Nikolaidou, Eds., ACM, pp. 8 :1–8 :10.
- [117] MARÍA POVEDA-VILLALÓN, R. G.-C. Saref extension for building, 2020.
- [118] MASCARDI, V., LOCORO, A., AND ROSSO, P. Automatic ontology matching via upper ontologies : A systematic evaluation. *IEEE Trans. Knowl. Data Eng.* 22, 5 (2010), 609–623.
- [119] MCGLINN, K., O’NEILL, E., GIBNEY, A., O’SULLIVAN, D., AND LEWIS, D. Simcon : A tool to support rapid evaluation of smart building application design using context simulation and virtual reality. *J. Univers. Comput. Sci.* 16, 15 (2010), 1992–2018.
- [120] MEIER, A., AND KAUFMANN, M. Nosql databases. In *SQL & NoSQL databases*. Springer, 2019, pp. 201–218.
- [121] METZLER, D. Using gradient descent to optimize language modeling smoothing parameters. In *SIGIR 2007 : Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007* (2007), W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, Eds., ACM, pp. 687–688.
- [122] MISHRA, A., AND JAIN, S. K. A survey on question answering systems with classification. *Journal of King Saud University-Computer and Information Sciences* 28, 3 (2016), 345–361.
- [123] MOKBEL, M. F., XIONG, X., HAMMAD, M. A., AND AREF, W. G. Continuous query processing of spatio-temporal data streams in place. *GeoInformatica* 9, 4 (2005), 343–365.
- [124] MONTEIRO, S., AND VIJAYA, B. Ontology population from complex sentences document. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)* (2017), pp. 951–954.
- [125] MUNIR, K., AND ANJUM, M. S. The use of ontologies for effective knowledge modelling and information retrieval. *Applied Computing and Informatics* 14, 2 (2018), 116–126.
- [126] MUNIR, K., ODEH, M., AND MCCLATCHEY, R. Ontology-driven relational query formulation using the semantic and assertional capabilities of OWL-DL. *Knowl. Based Syst.* 35 (2012), 144–159.
- [127] MUNIR, K., ODEH, M., MCCLATCHEY, R., KHAN, S., AND HABIB, I. Semantic information retrieval from distributed heterogeneous data sources. *CoRR abs/0707.0745* (2007).
- [128] MUSTAFA, J., KHAN, S., AND LATIEF, K. Ontology based semantic information retrieval. In *2008 4th International IEEE Conference Intelligent Systems* (2008), vol. 3, IEEE, pp. 22–14.

- [129] NAKANISHI, T., ZETTSU, K., KIDAWARA, Y., AND KIYOKI, Y. A context dependent dynamic interconnection method of heterogeneous knowledge bases by interrelation management function. In *Information Modelling and Knowledge Bases XXI, 19th European-Japanese Conference on Information Modelling and Knowledge Bases (EJC 2009), Maribor, Slovenia, June 1-5, 2009* (2009), T. Welzer, H. Jaakkola, Y. Kiyoki, T. Tokuda, and N. Yoshida, Eds., vol. 206 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, pp. 208–225.
- [130] NAMIOT, D. Time series databases. In *Selected Papers of the XVII International Conference on Data Analytics and Management in Data Intensive Domains (DAM-DID/RCDL 2015), Obninsk, Russia, October 13-16, 2015* (2015), L. A. Kalinichenko and S. Starkov, Eds., vol. 1536 of *CEUR Workshop Proceedings*, CEUR-WS.org, pp. 132–137.
- [131] NAQVI, S. N. Z., YFANTIDOU, S., AND ZIMÁNYI, E. Time series databases and influxdb. *Studienarbeit, Université Libre de Bruxelles 12* (2017).
- [132] NASAR, M., AND KAUSAR, M. A. Suitability of influxdb database for iot applications. *International Journal of Innovative Technology and Exploring Engineering* 8, 10 (2019), 1850–1857.
- [133] NEUMANN, G., AND SACALEANU, B. Part iv-multiple language question answering-experiments on robust nl question interpretation and multi-layered document annotation for a cross-language question/answering system. *Lecture Notes in Computer Science 3491* (2005), 411–422.
- [134] NILES, I., AND PEASE, A. Towards a standard upper ontology. In *2nd International Conference on Formal Ontology in Information Systems, FOIS 2001, Ogunquit, Maine, USA, October 17-19, 2001, Proceedings* (2001), ACM, pp. 2–9.
- [135] NITTEL, S. Real-time sensor data streams. *SIGSPATIAL Special* 7, 2 (2015), 22–28.
- [136] NURDIN, A., AND MAULIDEVI, N. 5w1h information extraction with cnn-bidirectional lstm. In *Journal of Physics : Conference Series* (2018), vol. 978, IOP Publishing, p. 012078.
- [137] OTERO-CERDEIRA, L., RODRÍGUEZ-MARTÍNEZ, F. J., AND GÓMEZ-RODRÍGUEZ, A. Ontology matching : A literature review. *Expert Syst. Appl.* 42, 2 (2015), 949–971.
- [138] PANDOLFO, L., PULINA, L., AND ADORNI, G. A framework for automatic population of ontology-based digital libraries. In *AI*IA 2016 : Advances in Artificial Intelligence - XVth International Conference of the Italian Association for Artificial Intelligence, Genova, Italy, November 29 - December 1, 2016, Proceedings* (2016), G. Adorni, S. Cagnoni, M. Gori, and M. Maratea, Eds., vol. 10037 of *Lecture Notes in Computer Science*, Springer, pp. 406–417.
- [139] PARHI, M., ACHARYA, B. M., AND PUTHAL, B. An effective mechanism to discover sensor web registry services for wireless sensor network under x-soa approach. In *Trendz in Information Sciences Computing(TISC2010)* (2010), pp. 197–201.

- [140] PARK, D., BANG, H., PYO, C. S., AND KANG, S. Semantic open iot service platform technology. In *IEEE World Forum on Internet of Things, WF-IoT 2014, Seoul, South Korea, March 6-8, 2014* (2014), IEEE Computer Society, pp. 85–88.
- [141] PETASIS, G., KARKALETSIS, V., PALIOURAS, G., KRITHARA, A., AND ZAVITSANOS, E. Ontology population and enrichment : State of the art. In *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution - Bridging the Semantic Gap*, G. Paliouras, C. D. Spyropoulos, and G. Tsatsaronis, Eds., vol. 6050 of *Lecture Notes in Computer Science*. Springer, 2011, pp. 134–166.
- [142] PETRAKIS, E. G., SOTIRIADIS, S., SOULTANOPOULOS, T., RENTA, P. T., BUYYA, R., AND BESSIS, N. Internet of things as a service (itaas) : Challenges and solutions for management of sensor data on the cloud and the fog. *Internet of Things 3* (2018), 156–174.
- [143] PORTER, M. E., AND HEPELMANN, J. E. How smart, connected products are transforming competition. *Harvard business review* 92, 11 (2014), 64–88.
- [144] RADEV, D. R., QI, H., WU, H., AND FAN, W. Evaluating web-based question answering systems. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain* (2002), European Language Resources Association.
- [145] RAHMAWATI, D., AND KHODRA, M. L. Word2vec semantic representation in multi-label classification for indonesian news article. In *2016 International Conference On Advanced Informatics : Concepts, Theory And Application (ICAICTA)* (2016), IEEE, pp. 1–6.
- [146] REPPLINGER, J. G.G. chowdhury. *Introduction to Modern Information Retrieval*. 3rd ed. london : Facet, 2010. 508p. alk. paper, \$90 (ISBN 9781555707156). LC2010-013746. *Coll. Res. Libr.* 72, 2 (2011), 194–195.
- [147] RICHARDSON, R., AND SMEATON, A. F. Using wordnet in a knowledge-based approach to information retrieval. *Dublin City University, School of Computer Applications : Dublin, Ireland* (1995).
- [148] RINALDI, A. M. An ontology-driven approach for semantic information retrieval on the web. *ACM Trans. Internet Techn.* 9, 3 (2009), 10 :1–10 :24.
- [149] RINALDI, A. M., AND RUSSO, C. A matching framework for multimedia data integration using semantics and ontologies. In *12th IEEE International Conference on Semantic Computing, ICSC 2018, Laguna Hills, CA, USA, January 31 - February 2, 2018* (2018), IEEE Computer Society, pp. 363–368.
- [150] RUIZ-MARTINEZ, J. M., MINARRO-GIMÉNEZ, J. A., CASTELLANOS-NIEVES, D., GARCIA-SÁNCHEZ, F., AND VALENCIA-GARCIA, R. Ontology population : an application for the e-tourism domain. *International Journal of Innovative Computing, Information and Control (IJICIC)* 7, 11 (2011), 6115–6134.
- [151] RUIZ-MARTÍNEZ, J. M., VALENCIA-GARCÍA, R., MARTÍNEZ-BÉJAR, R., AND HOFFMANN, A. G. Bioontology : A top level ontology based framework to populate biomedical ontologies from texts. *Knowl. Based Syst.* 36 (2012), 68–80.

- [152] RUSSELL, S. J., AND NORVIG, P. Artificial intelligence : a modern approach. malaysia, 2016.
- [153] RUSSOMANNO, D. J., KOTHARI, C., AND THOMAS, O. Sensor ontologies : from shallow to deep models. In *Proceedings of the Thirty-Seventh Southeastern Symposium on System Theory, 2005. SSST'05.* (2005), IEEE, pp. 107–112.
- [154] SABOU, M., D'AQUIN, M., AND MOTTA, E. Exploring the semantic web as background knowledge for ontology matching. *J. Data Semant.* 11 (2008), 156–190.
- [155] SALTON, G., FOX, E. A., AND WU, H. Extended boolean information retrieval. *Communications of the ACM* 26, 11 (1983), 1022–1036.
- [156] SCHARFFE, F., LIU, Y., AND ZHOU, C. Rdf-ai : an architecture for rdf datasets matching, fusion and interlink. In *Proc. IJCAI 2009 workshop on Identity, reference, and knowledge representation (IR-KR), Pasadena (CA US)* (2009), p. 23.
- [157] SHARMA, S., KUMAR, R., BHADANA, P., AND GUPTA, S. News event extraction using 5w1h approach & its analysis. *International Journal of Scientific & Engineering Research* 4, 5 (2013), 2064–2068.
- [158] SHAW, R., TRONCY, R., AND HARDMAN, L. LODÉ : linking open descriptions of events. In *The Semantic Web, Fourth Asian Conference, ASWC 2009, Shanghai, China, December 6-9, 2009. Proceedings* (2009), A. Gómez-Pérez, Y. Yu, and Y. Ding, Eds., vol. 5926 of *Lecture Notes in Computer Science*, Springer, pp. 153–167.
- [159] SHETH, A. P., HENSON, C. A., AND SAHOO, S. S. Semantic sensor web. *IEEE Internet Comput.* 12, 4 (2008), 78–83.
- [160] SINGHAL, A. Modern information retrieval : A brief overview. *IEEE Data Eng. Bull.* 24, 4 (2001), 35–43.
- [161] SOYLU, A., GIESE, M., JIMÉNEZ-RUIZ, E., KHARLAMOV, E., ZHELEZNYAKOV, D., AND HORROCKS, I. Optiquevqs : towards an ontology-based visual query system for big data. In *Fifth International Conference on Management of Emergent Digital EcoSystems, MEDES '13, Luxembourg, Luxembourg, October 29-31, 2013* (2013), L. Ladid, A. Montes, P. A. Bruck, F. Ferri, and R. Chbeir, Eds., ACM, pp. 119–126.
- [162] SUCHANEK, F. M., IFRIM, G., AND WEIKUM, G. LEILA : learning to extract information by linguistic analysis. In *Proceedings of the 2nd Workshop on Ontology Learning and Population : Bridging the Gap between Text and Knowledge@COLING/ACL 2006, Sydney, Australia, July 22, 2006* (2006), P. Buitelaar, P. Cimiano, and B. Loos, Eds., Association for Computational Linguistics, pp. 18–25.
- [163] SUKANYA, C., GOKUL, R., AND PAUL, V. A survey on object recognition methods. *International Journal of Science, Engineering and Computer Technology* 6, 1 (2016), 48.
- [164] SUN, J., AND ZHOU, J. Querying sensor networks with extended SQL. In *Proceedings of the IEEE International Conference on Wireless Communications, Networking and Information Security, WCNIS 2010, 25-27 June 2010, Beijing, China* (2010), IEEE, pp. 634–638.

- [165] TANEV, H., AND MAGNINI, B. Weakly supervised approaches for ontology population. In *Ontology Learning and Population : Bridging the Gap between Text and Knowledge*, P. Buitelaar and P. Cimiano, Eds., vol. 167 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2008, pp. 129–143.
- [166] TANG, M., CHEN, J., AND CHEN, H. Semoir : An ontology-based semantic information retrieval system. In *2020 IEEE 20th International Conference on Software Quality, Reliability and Security Companion (QRS-C) (2020)*, IEEE, pp. 204–208.
- [167] THANTRIWATTE, T., AND KEPPETIYAGAMA, C. Nosql query processing system for wireless ad-hoc and sensor networks. In *2011 International Conference on Advances in ICT for Emerging Regions (ICTer) (2011)*, IEEE, pp. 78–82.
- [168] THIÉBLIN, É., HAEMMERLÉ, O., HERNANDEZ, N., AND TROJAHN, C. Survey on complex ontology matching. *Semantic Web 11*, 4 (2020), 689–727.
- [169] THOMA, M., MEYER, S., SPERNER, K., MEISSNER, S., AND BRAUN, T. On iot-services : Survey, classification and enterprise integration. In *2012 IEEE International Conference on Green Computing and Communications (2012)*, IEEE.
- [170] TIWARI, S. *Professional nosql*. John Wiley & Sons, 2011.
- [171] TSIFTES, N., AND DUNKELS, A. A database in every sensor. In *Proceedings of the 9th International Conference on Embedded Networked Sensor Systems, SenSys 2011, Seattle, WA, USA, November 1-4, 2011 (2011)*, J. Liu, P. A. Levis, and K. Römer, Eds., ACM, pp. 316–332.
- [172] TURTLE, H., AND CROFT, W. B. Inference networks for document retrieval. In *Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval (1989)*, pp. 1–24.
- [173] UTHAYAN, K., AND ANANDHA MALA, G. Hybrid ontology for semantic information retrieval model using keyword matching indexing system. *The Scientific World Journal 2015 (2015)*.
- [174] VALLET, D., FERNÁNDEZ, M., AND CASTELLS, P. An ontology-based information retrieval model. In *The Semantic Web : Research and Applications, Second European Semantic Web Conference, ESWC 2005, Heraklion, Crete, Greece, May 29 - June 1, 2005, Proceedings (2005)*, A. Gómez-Pérez and J. Euzenat, Eds., vol. 3532 of *Lecture Notes in Computer Science*, Springer, pp. 455–470.
- [175] VAN DER VEEN, J. S., VAN DER WAAIJ, B., AND MEIJER, R. J. Sensor data storage performance : SQL or nosql, physical or virtual. In *2012 IEEE Fifth International Conference on Cloud Computing, Honolulu, HI, USA, June 24-29, 2012 (2012)*, R. Chang, Ed., IEEE Computer Society, pp. 431–438.
- [176] VAN HAGE, W. R., MALAISÉ, V., SEGERS, R., HOLLINK, L., AND SCHREIBER, G. Design and use of the simple event model (SEM). *J. Web Semant.* 9, 2 (2011), 128–136.
- [177] VARGAS, S. Novelty and diversity enhancement and evaluation in recommender systems and information retrieval. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast*

- , QLD, Australia - July 06 - 11, 2014 (2014), S. Geva, A. Trotman, P. Bruza, C. L. A. Clarke, and K. Järvelin, Eds., ACM, p. 1281.
- [178] VENKATA KRISHNA MOHAN SUNKARA, ASHOK SAMAL, L.-K. S. *A DATA DRIVEN APPROACH TO IDENTIFY JOURNALISTIC 5WS FROM TEXT DOCUMENTS*. PhD thesis, 2019.
- [179] VERMESAN, O., AND FRIESS, P. *Internet of things : converging technologies for smart environments and integrated ecosystems*. River publishers, 2013.
- [180] VIJAYARAJAN, V., DINAKARAN, M., TEJASWIN, P., AND LOHANI, M. A generic framework for ontology-based information retrieval and image retrieval in web data. *Hum. centric Comput. Inf. Sci.* 6 (2016), 18.
- [181] VILLCA-ROCHA, A., ZHENG, M., DUAN, C., AND WANG, H. Towards semantic search in building sensor data. In *BuildSys '21 : The 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, Coimbra, Portugal, November 17 - 18, 2021* (2021), X. F. Jiang, O. Gnawali, and Z. Nagy, Eds., ACM, pp. 164–167.
- [182] VIZZINI, J. *Data interlinking with relational concept analysis*. PhD thesis, Université Grenoble Alpes, 2017.
- [183] VOLZ, J., BIZER, C., GAEDKE, M., AND KOBILAROV, G. Silk - A link discovery framework for the web of data. In *Proceedings of the WWW2009 Workshop on Linked Data on the Web, LDOW 2009, Madrid, Spain, April 20, 2009* (2009), C. Bizer, T. Heath, T. Berners-Lee, and K. Idehen, Eds., vol. 538 of *CEUR Workshop Proceedings*, CEUR-WS.org.
- [184] WANG, C., HUANG, X., QIAO, J., JIANG, T., RUI, L., ZHANG, J., KANG, R., FEINAUER, J., MCGRAIL, K. A., WANG, P., ET AL. Apache iotdb : time-series database for internet of things. *Proceedings of the VLDB Endowment* 13, 12 (2020), 2901–2904.
- [185] WANG, M. A survey of answer extraction techniques in factoid question answering. *Computational Linguistics* 1, 1 (2006), 1–14.
- [186] WANG, S., YU, M., GUO, X., WANG, Z., KLINGER, T., ZHANG, W., CHANG, S., TESAURO, G., ZHOU, B., AND JIANG, J. R³ : Reinforced ranker-reader for open-domain question answering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018* (2018), S. A. McIlraith and K. Q. Weinberger, Eds., AAAI Press, pp. 5981–5988.
- [187] WANG, W., AND BARNAGHI, P. M. Semantic annotation and reasoning for sensor data. In *Smart Sensing and Context, 4th European Conference, EuroSSC 2009, Guildford, UK, September 16-18, 2009. Proceedings* (2009), P. M. Barnaghi, K. Moessner, M. Presser, and S. Meissner, Eds., vol. 5741 of *Lecture Notes in Computer Science*, Springer, pp. 66–76.

- [188] WANG, W., ZHAO, D., AND WANG, D. Chinese news event 5w1h elements extraction using semantic role labeling. In *2010 Third International Symposium on Information Processing* (2010), IEEE, pp. 484–489.
- [189] WANG, W., ZHAO, D., ZOU, L., WANG, D., AND ZHENG, W. Extracting 5w1h event semantic elements from chinese online news. In *International Conference on Web-Age Information Management* (2010), Springer.
- [190] WANG, W., ZHAO, D., ZOU, L., WANG, D., AND ZHENG, W. Extracting 5w1h event semantic elements from chinese online news. In *Web-Age Information Management, 11th International Conference, WAIM 2010, Jiuzhaigou, China, July 15-17, 2010. Proceedings* (2010), L. Chen, C. Tang, J. Yang, and Y. Gao, Eds., vol. 6184 of *Lecture Notes in Computer Science*, Springer, pp. 644–655.
- [191] WANG, X. Deep learning in object recognition, detection, and segmentation. *Found. Trends Signal Process.* 8, 4 (2016), 217–382.
- [192] WANG, Y., LIU, W., AND BELL, D. A. A concept hierarchy based ontology mapping approach. In *Knowledge Science, Engineering and Management, 4th International Conference, KSEM 2010, Belfast, Northern Ireland, UK, September 1-3, 2010. Proceedings* (2010), Y. Bi and M. Williams, Eds., vol. 6291 of *Lecture Notes in Computer Science*, Springer, pp. 101–113.
- [193] WEN, K., LI, R., AND LI, B. Searching concepts and association relationships based on domain ontology. In *GCC 2010, The Ninth International Conference on Grid and Cloud Computing, Nanjing, Jiangsu, China, 1-5 November 2010* (2010), IEEE Computer Society, pp. 432–437.
- [194] WÖLGER, S., SIORPAES, K., BÜRGER, T., SIMPERL, E., THALER, S., AND HOFER, C. A survey on data interlinking methods. *STI TECHNICAL REPORT* (2011).
- [195] WU, H., CHELMIS, C., SORATHIA, V., ZHANG, Y., PATRI, O. P., AND PRASANNA, V. K. Enriching employee ontology for enterprises with knowledge discovery from social networks. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (2013), pp. 1315–1322.
- [196] YAMAN, B., PASIN, M., AND FREUDENBERG, M. Interlinking scigraph and dbpedia datasets using link discovery and named entity recognition techniques. In *2nd Conference on Language, Data and Knowledge, LDK 2019, May 20-23, 2019, Leipzig, Germany* (2019), M. Eskevich, G. de Melo, C. Fäth, J. P. McCrae, P. Buitelaar, C. Chiarcos, B. Klimek, and M. Dojchinovski, Eds., vol. 70 of *OASICS*, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, pp. 15 :1–15 :8.
- [197] YAMAN, S., HAKKANI-TÜR, D., AND TÜR, G. Combining semantic and syntactic information sources for 5-w question answering. In *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009* (2009), ISCA, pp. 2707–2710.
- [198] YAMAN, S., HAKKANI-TÜR, D., TUR, G., GRISHMAN, R., HARPER, M., MCKEOWN, K. R., MEYERS, A., AND SHARMA, K. Classification-based strategies for combining

- multiple 5-w question answering systems. In *Tenth Annual Conference of the International Speech Communication Association* (2009).
- [199] YANG, Y., CAO, Q., AND JIANG, H. Edgedb : An efficient time-series database for edge computing. *IEEE Access* 7 (2019), 142295–142307.
- [200] ZHANG, G.-Q., SIEGLER, T., SAXMAN, P., SANDBERG, N., MUELLER, R., JOHNSON, N., HUNSCHER, D., AND ARABANDI, S. Visage : a query interface for clinical research. *Summit on translational bioinformatics 2010* (2010), 76.
- [201] ZHANG, H., GUO, Y., LI, Q., GEORGE, T. J., SHENKMAN, E., MODAVE, F., AND BIAN, J. An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. *BMC Medical Informatics Decis. Mak.* 18, S-2 (2018), 129–147.
- [202] ZHANG, J., DENG, B., AND LI, X. Concept based query expansion using wordnet. In *2009 International e-Conference on Advanced Science and Technology* (2009), IEEE, pp. 52–55.
- [203] ZHANG, W., YOSHIDA, T., AND TANG, X. Text classification based on multi-word with support vector machine. *Knowledge-Based Systems* 21, 8 (2008), 879–886.
- [204] ZHAO, F., FANG, F., YAN, F., JIN, H., AND ZHANG, Q. Expanding approach to information retrieval using semantic similarity analysis based on wordnet and wikipedia. *Int. J. Softw. Eng. Knowl. Eng.* 22, 2 (2012), 305–322.
- [205] ZHAO, L., ICHISE, R., MITA, S., AND SASAKI, Y. Core ontologies for safe autonomous driving. In *International Semantic Web Conference (Posters & Demos)* (2015).
- [206] ZHU, F., TURNER, M., KOTSIPOPOULOS, I., BENNETT, K. H., RUSSELL, M., BUDGEN, D., BRERETON, P., KEANE, J. A., LAYZELL, P. J., RIGBY, M., AND XU, J. Dynamic data integration using web services. In *Proceedings of the IEEE International Conference on Web Services (ICWS'04), June 6-9, 2004, San Diego, California, USA* (2004), IEEE Computer Society, pp. 262–269.
- [207] ZOU, L., ÖZSU, M. T., CHEN, L., SHEN, X., HUANG, R., AND ZHAO, D. gstore : a graph-based sparql query engine. *The VLDB journal* 23, 4 (2014), 565–590.