Ontology-Based Integration of Occupational Health Data: Method and Case Studies

Cassandra Barbey
Department of Pollutant Metrology, INRS
Univ Rennes, Inserm, EHESP, Irset- UMR_S 1085
Vandœuvre-lès-Nancy, France
Email: cassandra.barbey@inrs.fr

Malika Smaïl-Tabbone
LORIA
Université de Lorraine, CNRS, LORIA, UMR 7503,
Vandœuvre-lès-Nancy, France
Email: malika.smail@loria.fr

Nathalie Bonvallot
Univ Rennes, Inserm, EHESP, Irset - UMR_S 1085
Rennes, France
Email: nathalie.bonvallot@ehesp.fr

Frédéric Clerc
Department of Pollutant Metrology, INRS
Vandœuvre-lès-Nancy, France
Email: frederic.clerc@inrs.fr

***Abstract*— The data related to occupational health exhibit diverse characteristics and are not inherently designed to interoperate; however, they contain complementary information. The integration of such data has the potential to enhance the current understanding of occupational health risks. Therefore, the objective of this study is to analyse heterogeneous data derived from 10 French occupational databases provided by 6 French institutes. An Ontology-Based Data Integration (OBDI) approach was employed, involving the mapping of data sources to a domain-specific ontology, namely the Occupational Exposure Ontology (OExO). In addition to OExO, four other ontologies were utilised: the Occupational Exposure Thesaurus (TEP) for occupational nuisances or hazards, the International Classification of Diseases (ICD-10) for medical conditions, the French Nomenclature of Activities (NAF) for industry sectors, and the Professions and Socio-professional Categories (PCS) for occupational classifications. The integration of these data is primarily achieved through the concept of the "occupational group", defined as a cohort of individuals of the same gender, engaged in the same occupation, and employed within the same industry sector. The study presents two case studies derived from the integrated knowledge base: a quantitative analysis identifying occupational groups with the highest exposure to nuisances and disease prevalence, and a qualitative analysis evaluating the consistency of information associated with each nuisance and disease.**

Keywords*— *ontologies; integration data; heterogeneous data; occupational health.

I. INTRODUCTION

Workers are exposed to several occupational nuisances that can have an effect on their safety or health and lead to occupational accidents or diseases. Moreover, interactions between these nuisances can affect health differently, reducing the effectiveness of risk mitigation measures often designed for single nuisances. The implementation of relevant preventive actions requires knowledge about these interactions, which is still limited.

In France, several national organisations collect occupational nuisance data and health data from surveys, declarations, or surveillance systems. These databases have different characteristics (objectives, collection method, target population, etc.) and provide much information but they were not designed to be used jointly. Some databases have been created to be representative of the population of French workers, while others exhibit more restricted scope. The information they contain may also be different, in line with their initial objective (to monitor, reference, describe, encourage, analyse, group together, etc.).

Analytical methodologies whose aim is to use occupational health data from several databases related to occupational risks prevention together have been identified. Some studies focus on a specific subject, such as that of L. Rollin et al. [1] about the occupational diseases faced by women in the homecare sector. Others are part of a larger project, such as Datamining project [2], in which both administrative recorded data and data from surveys are used. Following the same path of integrating health-related data related to the surveillance of elderly people, Dandan et al. [3] attempted to formalise the knowledge using ontologies for integrating data from sensors, surveys and personal health records. The DataPOST project is part of this trend, with the aim of developing a methodology for extracting knowledge about occupational nuisances and health outcomes, while relying on an ontological approach in order to bring together information from ten databases.

For this purpose, the data are integrated using an Ontology-Based Data Integration (OBDI) approach [4], and used to qualify and quantify multiple nuisances and health effects. The statistical unit used for analysis is named “occupational group”, which is a set of individuals of the same sex sharing the same occupation and working in the same sector of activity. The variables used for the definition of an occupational group are defined in all databases. These integrated data are then used in various analyses. We selected two examples that will be presented in Sections 5 and 6:

- A quantitative analysis to measure the degree of exposure of occupational groups to several nuisances and diseases by creating relevant indicators.
- A qualitative analysis to check the consistency of the data on each nuisance and disease provided by the databases and the validity of the relevant indicators constructed.

The rest of the article is structured into 5 sections: the introduction is followed by Sections 2 and 3, which define OBDI and OExO; Section 4 details the general representation of the data using Sections 2 and 3. Sections 5 and 6 present the case studies, detailing the methodology and results. Section 7 concludes and presents the outlook for the future.

II. ONTOLOGY-BASED DATA INTEGRATION (OBDI) APPROACH

The ODBI approach aims at integrating heterogeneous data by leveraging on an ontology that contains a semantic description of the concepts and their relationships in a domain of interest. This approach is built around three elements: the data sources, a heterogeneous repository where data are stored; the domain ontology, a formal description of that domain made by the organisations involved; and the mapping between them acting as the reconciliation structure (Figure 1).

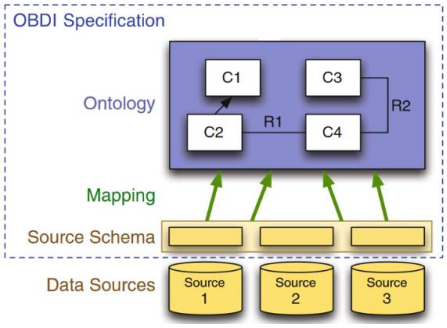

Figure 1. Ontology-based data integration adapted from Calvanese et al. (2017) [4].

This approach will be our methodological starting point. The context of our study is more complex, and several ontologies are required to integrate the available data and their relationships. It is necessary to have an integrated representation of data and ontologies similarly to the Occupational Exposure Ontology (OExO) [5].

III. OCCUPATIONAL EXPOSURE ONTOLOGY (OEXO)

The central knowledge representation used for this study relies on OExO, which is itself an extension of the Exposure science Ontology [6]. OExO consists of four central nuisance concepts: receptor, stressor, event and outcome; each of which is described by several child terms and attributes. The receptor is an individual worker or a population of workers that may be exposed to a stressor. The stressor represents an agent, activity or event that can affect the nuisance receptor, a chemical substance for example. The interaction between the two is called an exposure event that can lead to a health outcome, a disease for example (Figure 2).

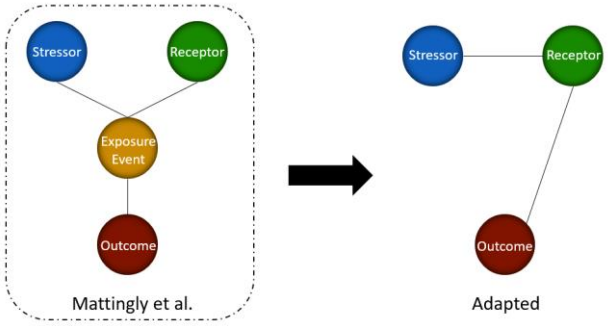

Figure 2. Main concepts of the exposure science ontology from Mattingly et al. (2012) [6] and their adaptation to our context

In this work, the receptor corresponds to the occupational group and the stressor to the occupational nuisance. In this work, unlike to OExO, the exposure event concept is not defined and is therefore not used. However, the available information in our databases allows us to link the occupational group the outcome.

IV. GENERAL REPRESENTATION OF THE HEALTH OCCUPATIONAL DATA USING ODBI AND ADAPTED OEXO

In order to integrate health occupational data using ODBI and adapted OExO main concepts, we grouped together several ontologies and related them to the available data, in the form of a knowledge base formalised as a conceptual data model (Figure 3). The following paragraphs will describe the ontologies we used for the Stressor/Receptor/Outcome concepts, the data sources, and the mappings we had to define between the latter and the former.

A. Ontologies

Four ontologies are used:

- The Occupational Exposure Thesaurus (TEP) [7], used to characterise and group the nuisances. This is a reference system designed in 2014 by the French agency for health safety to collect uniformly data on occupational nuisances. TEP is organised in 8 hierarchical levels representing around 8,300 nuisance concepts.
- ICD-10 is the tenth revision of the international classification of diseases [8], an international compilation on the causes and consequences of human disease designed and maintained by the World Health Organisation. It provides a common health language by using around 150 000 codify clinical terms [9]. It consists of 22 chapters subdivided into several blocks of three-character categories which can also be subdivided into four-character subcategories.
- The statistical classification of economic activities in the European Community is used to organise the information about economic and social activities. In this case, its French version named NAF [10] is used for the

definition of the occupational group. It is divided into 5 nested levels.

- The Professions and Socio-professional Categories (PCS) [11] results from a statistical classification conducted by the National Institute of Statistics and Economic Studies that brings together occupations from the same social background. This ontology is used for the definition of the occupational groups. It is divided into 4 nested levels of job designations.

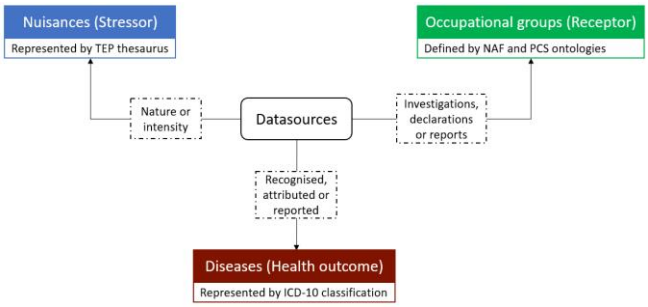

Figure 3. General representation of used data and ontologies, linked to the adapted OExO ontology.

B. *Data sources and their schema*

Ten data sources are used. Six of them provide information on occupational nuisances: SUMER [1], C2P [12], COLCHIC and SCOLA [13], COLPHY [14] and MatGene [15]; one data source concerns occupational diseases: AT-MP [16]; and three on both: Evrest [1], MCP [1] and RNV3P [1]. These data sources are described in Table I.

TABLE I. TABLE DESCRIBING THE DIFFERENT DATA SOURCES AND THE INFORMATION THEY CONTAIN

Data source	Collection method	Original statistical unit	Content	Example
Sumer	National surveys	Worker	340 columns representing nuisances to which workers are exposed to.	Nuisance to lead [yes ; no]
C2P	Regulatory Declarations	Worker	10 columns representing nuisances to which employers declared the worker are exposed to.	Nuisance to repetitive movements [yes ; no]
Colchic / Scola	Sampling and analysis of workplace air by specialised chemistry laboratories	Measurement	460 columns representing the measurement of the Intensity of the concentration of 230 substances in the air with regards to the regulatory limit value.	Lead Intensity [moderate ; high ; very high]
Colphy	Historical measurements and sampling	Measurement	4 columns representing the measurement of the intensity of the emissivity of 2 physical nuisances with regards to the regulatory limit value.	Whole body vibration Intensity [moderate ; high ; very high]
MatGene	Historical and census information	Occupational group	4 columns representing the nature and intensity of nuisance according to occupation.	Night work [yes ; no]
AT-MP	Medical consultation	Worker	86 columns representing occupational recognised diseases among workers.	Spondylopathies [yes ; no]
Evrest	Systematic occupational health interviews	Worker	45 columns representing the percentage of workers concerns by nuisance. 15 columns representing clinical signs. 15 columns representing first treatment.	Noise [yes ; no] Treatment for hearing problems [yes ; no]
MCP	Compulsory professional medical consultation	Worker	720 columns representing nuisances associated with reported occupational diseases and 56 columns representing these work-related diseases.	[Allergic contact dermatitis] [Chemical agents]
RNV3P	Medical consultation with a specialist of CCPPE	Health problem	129 columns representing the occupational diseases identified and 908 columns representing the nuisances probably linked to these diseases.	[Scoliosis] [Heavy loads ; Awkward postures]

C. *Mapping between data schemas and ontologies*

The mapping between data and ontologies is a 3-stage process:

- Mapping to define “occupational groups”: All combinations of variables relating to sector of activity (NAF), occupation (PCS) and sex are created.
- Mapping to standardise “health outcomes”: each variable in the AT-MP, MCP and RNV3P data are associated to the disease codes present in the ICD-10 classification.
- Mapping to define occupational nuisances: each nuisance variable in the data sources is associated to the nuisance it represents in the TEP. For example, the “extreme temperatures” variable in C2P, which refers to all temperatures below or equal to 5°C or at least equal to 30°C, will be linked to the “extreme thermal environment” nuisance in the TEP. This part was carried out on data from multiple sources (SUMER, MatGene, C2P, Evrest, COLCHIC/SCOLA, COLPHY, MCP, RNV3P).

The integrated data are then used in two case studies, the results of which are presented for the construction sector. 12,835 occupational groups were created, including 816 for the construction sector. Information is available for 308 nuisances and 174 diseases. The analyses were carried out using RStudio software, an integrated development environment (v.4.3.2) [17].

V. CASE STUDY ONE: QUANTITATIVE ANALYSIS

An indicator is created for each nuisance and disease to represent its importance for each occupational group.

A. *Indicator construction method for simple nuisance or disease*

The nature of information contained in databases can be:

• "Quantification": number of workers exposed to the nuisance (SUMER and MatGene).

• "Declarative": number of workers who declared (or have been declared by their employer) to be exposed to the nuisance (C2P and Evrest).

• "Intensity": number of intensity assessments to the nuisance recorded and maximum intensity of the nuisance (COLCHIC/SCOLA and COLPHY).

• "Plausibility": number of occupational nuisances which are the cause of worker's diseases as assessed by occupational physicians (MCP and RNV3P).

• "Disease": number of occupational diseases (AT-MP, MCP and RNV3P).

As the data are heterogeneous, the values stored for each type of information do not have the same scale (see Table 1 for an example). Therefore, the raw values were converted into non-parametric values. To achieve this, the original data source values were discretised into a scale of 1 to 10, with 1 representing 'very low' and 10 representing 'very high.' For each type of information, a score was computed: the arithmetic average of the discretised values divided by 10, ensuring the score value is between 0 and 1. The four nuisance scores were summed up to define a nuisance indicator. The disease score was used as is.

Figure 4 shows the main stages in the creation and construction of indicators.

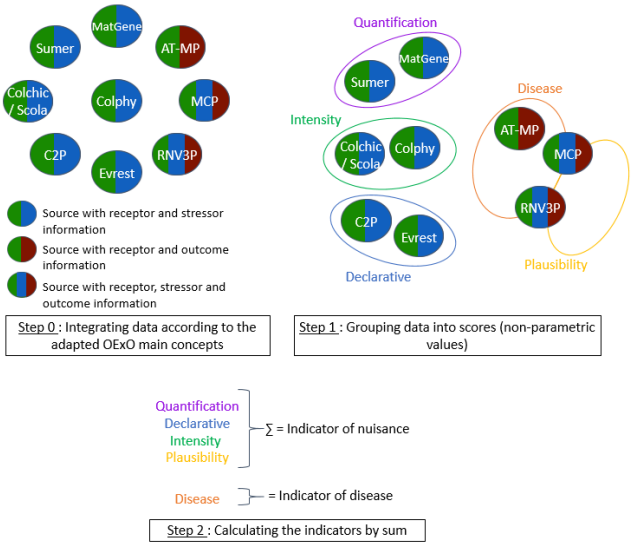

Figure 4. Main stages in the indicator construction method.

B. *Example: focus on the occupational group "43_632_1"*

The occupational group "43_632_1" represents skilled male workers in the "special construction" activity sector. This group is exposed to manual handling of heavy loads and suffering from arthrosis. The data confirms nuisance in the SUMER, C2P, MCP and RNV3P sources, as well as the presence of diseases in the AT-MP, MCP and RNV3P sources (Table II).

Indicator of nuisance "manual handling of heavy loads": The "quantification" score corresponds to the value present in the SUMER data source, MatGene having no information concerning this nuisance. The "declarative" score corresponds to the value present in the C2P data source, Evrest having no information concerning this nuisance. The "intensity" score is 0, COLCHIC/SCOLA and COLPHY having no information concerning this nuisance. The MCP and RNV3P data sources provide information; the "plausibility" score is calculated using both sources.

Indicator of disease "arthrosis": The AT-MP, MCP and RNV3P data sources provide information; the "disease" score is calculated using all three sources.

TABLE II. SUMMARY TABLE OF DATA ON MANUAL HANDLING OF HEAVY LOADS AND ARTHROSIS FOR THE OCCUPATIONAL GROUP "43_632_1"

	Data source	***Original data source value***	***Discretised value***	***Scores***	***Indicator***
Manual handling of heavy loads	Sumer	301,078.6	10	Quantification : 0.7	Nuisance indicator : 2.7
	MatGene	/	/		
	C2P	764.14	10	Declarative : 1	
	Evrest	/	/		
	Colchic /Scola	/	/	Intensity : 0	
	Colphy	/	/		
	MCP	321	10	Plausibility : 1	
	RNV3P	362	10		
Arthrosis	MCP	45	10	Disease : 10	Disease indicator : 10
	RNV3P	27	10		
	AT-MP	204	10		

C. *Heatmap: visualisation of nuisances and diseases indicators in the construction sector*

In the construction sector, 308 indicators for nuisances and 174 indicators for diseases were created. All the indicators are then represented in the form of a heatmap showing the occupational groups most at risk and most affected by diseases (Figure 5). The x-axis represents the various occupational groups, arranged in ascending order based on the number of exposures or diseases. Occupational groups with fewer exposures or diseases are positioned on the left, while those with a greater number of exposures or diseases are placed on the right. The y-axis displays the nuisances or diseases, ordered by the cumulative sum of their respective indicators. The upper portion of the y-axis corresponds to nuisances or diseases for which a large

number of occupational groups are exposed or affected, whereas the lower portion indicates nuisances or diseases associated with a smaller number of exposed or affected occupational groups.

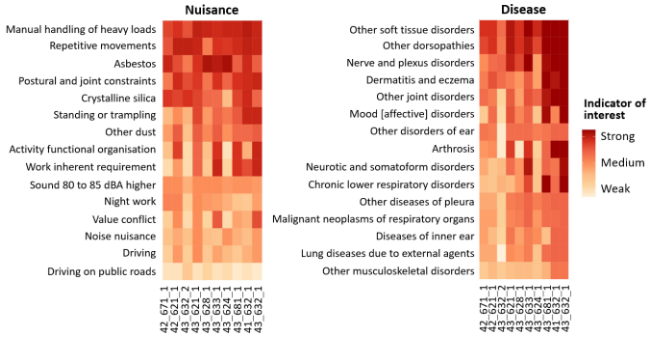

Figure 5. Heatmap of indicators by occupational group and by nuisance and disease: focus on 10 sickest occupational group and 15 nuisances and diseases.

For example, the occupational group "43_632_1" mentioned above is the most exposed and the most affected by occupational diseases because there is a lot of high valued indicators for nuisances and diseases. The most exposed occupational groups are particularly exposed to physical nuisances, such as postural and joint constraints, manual handling of heavy loads, repetitive movement, stranding and trampling, hand-arm vibration, as well as to certain chemical nuisances, such as asbestos and crystalline silica; nuisances typically expected in the construction sector. These groups are also heavily affected by musculoskeletal disorders like dorsopathies, nerve root and plexus disorders; diseases typically present in the construction sector. This work highlights the need to strengthen the preventive measures currently in place, for example by powered exoskeletons for reducing the risks related to the manual handling of heavy loads [18].

VI. CASE STUDY TWO: QUALITATIVE ANALYSIS

Our objective here is to measure the consistency of information contained in distinct databases in order to assess the complementarity of such sources. To do so, we created a consistency score for each nuisance and disease.

A. *Consistency score construction method*

This score is defined as the number of sources containing a value greater than 0 for a nuisance or a disease. The higher the number of sources, the greater the consistency between them.

B. *Example: focus on the occupational group "43_632_1"*

Following the above example on the occupational group "43_632_1", SUMER, C2P, MCP and RNV3P data sources confirm exposure to the nuisance "manual handling of heavy loads" with values greater than 0. The MatGene, COLCHIC/SCOLA and COLPHY data sources do not contain any information about this nuisance. They are therefore not taken into consideration when calculating the consistency score. The AT-MP, MCP and RNV3P data sources confirm the presence of disease "arthrosis" with also values greater than 0. The consistency scores for manual handling of heavy loads and for arthrosis will therefore be strong.

C. *Heatmap modelling of consistency score in the construction sector*

These scores are also represented in the form of a heatmap showing the groups for which the consistency of the information is strongest (Figure 6). The x-axis in this figure corresponds to that of the preceding heatmap. However, the y-axis differs slightly, as the values depicted here are derived from the consistency score rather than from the indicators. The upper portion of the y-axis indicates the nuisances or diseases for which the information is most coherent, while the lower portion reflects the nuisances or diseases associated with a smaller number of occupational groups demonstrating high consistency.

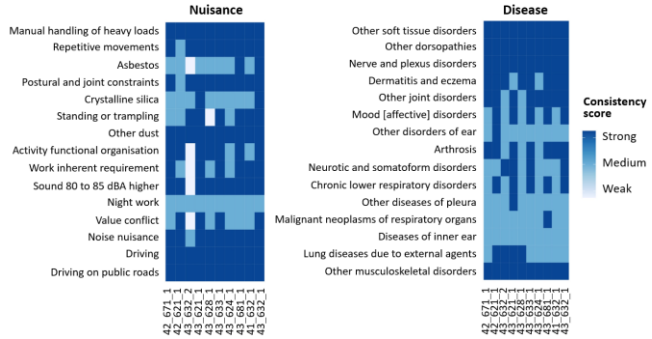

Figure 6. Heatmap of the consistency score by occupational group and by nuisance and disease: focus on 10 sickest occupational group and 15 nuisances and diseases.

For example, the "43_632_1" occupational group consistency scores for arthrosis and manual handling of heavy loads are high. We can also see that this group has a high representation of high scores for nuisances and a more moderate representation for diseases. The consistency scores are overall lower for diseases than for nuisances. This is due to the imbalance in the visibility of diseases. Some diseases that could be recognised as occupational are often not declared as such (hearing loss, for example). Information on these diseases are therefore not included in the main database on occupational diseases (AT-MP). However, these diseases may be attributed or reported in other databases. It highlights the complementarity of disease-related data sources and the need for integrating all data sources in order to get a broader picture.

VII. CONCLUSION AND FUTURE WORK

Structuring the data using the OBDI approach allowed us to implement two approaches to analyse the occupational health data. To the best of our knowledge, this study is the first attempt of integrating 10 heterogeneous data sources relying on four domain ontologies for the construction of indicators allows visualising information and highlighting occupational groups exposed to multiple nuisances and

victims of diseases. The consistency score is useful to check the validity of these indicators and the complementarity of the data.

The OExO ontology fits well with the concepts covered in the available data, although some modifications were applied to adapt it to our context. However, we could consider the reverse approach: adjust the data to the ontology. In this case, a data row in each source could be considered as a "nuisance event", such as defined in OExO. For this, we would need to explore further the ontology concepts related to nuisance events, in particular "assay" corresponding to all the features needed to assess the "nuisance event".

The two case studies presented here are examples of what can be done with data. We consider that our methodological proposal for data integration will enable us to integrate other data sources without any difficulties. Thus, a straightforward extension of the first use case would be to insert other data concerning the total number of workers per occupational group, in order to better assess the proportions of exposed and diseased workers.

The contextualisation of the data we propose opens new perspectives for their use and analyses for risk assessment purposes. For example, it would be possible to search for correlations between nuisance or co-nuisance indicators and disease indicators, in order to identify the main risks and subsequently create a tool for risk assessment [19]. Another possible perspective would be to use the indicators to establish worker nuisance profiles to centralise information on all possible nuisances or co-nuisances in a specific occupation [20]. These profiles could then be used to anticipate future occupational diseases and implement job-specific safety measures. Other work in the health sector is focused on the contextualisation of heterogeneous data to define a common semantic to facilitate knowledge sharing [21][22]. This generalisation would enable the development of responses adapted to current and future health concerns by means of tools and queries. It would also open the door to interdisciplinarity, and provide knowledge based on less theoretical situations.

We would like to generalise our methodology so that it can accommodate other data sources and subsequently facilitate data sharing. To achieve this, improvements are envisaged, notably in the mapping between data and ontologies, which remains complex, particularly regarding the occupations.

ACKNOWLEDGEMENT

The authors thank E. Algava and M. Duval (Dares), L. Meunier (CNAM), C. Pilorget, J. Chatelot and J. Homère (SPF), C. Nisse and A. Aachimi (Anses), L. Rollin and A. Leroyer (Evrest) for providing the data.

REFERENCES

[1] L. Rollin et al., "Complementarity of 4 data bases in occupational health", *Arch. Mal. Prof. Environ.*, vol. 82, no. 3, pp. 261-276, May 2021, doi: 10.1016/j.admp.2020.11.002.

[2] https://data.risquesautravail.be/fr, [last accessed Sept., 2024].

[3] R. Dandan, S. Despres, and J. Nobecourt, "OAFE: An Ontology for the Description of Elderly Activities", in *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, Nov. 2018, pp. 396-403. doi: 10.1109/SITIS.2018.00068.

[4] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati, "Ontology-Based Data Access and Integration", in *Encyclopedia of Database Systems*, 2017, pp. 1-7. doi: 10.1007/978-1-4899-7993-3_80667-1.

[5] D. A. Vallero, "A Draft Ontology for Occupational Exposure", Oct. 2016, doi: 10.13140/RG.2.2.16261.14564.

[6] C. J. Mattingly, T. E. McKone, M. A. Callahan, J. A. Blake, and E. A. C. Hubal, "Providing the Missing Link: the Exposure Science Ontology ExO", *Environ. Sci. Technol.*, vol. 46, no. 6, pp. 3046-3053, Mar. 2012, doi: 10.1021/es2033857.

[7] J. Bloch et al., "National Network for Monitoring Prevention of an Occupational Disease (RNV3P) - 2022 Annual Report", Oct. 2023.

[8] https://www.who.int, [last accessed Sept. 2024].

[9] J. A. Hirsch et al., "ICD-10: History and Context", *Am. J. Neuroradiol.*, vol. 37, no. 4, pp. 596-599, Apr. 2016, doi: 10.3174/ajnr.A4696.

[10] https://www.insee.fr, [last accessed Sept. 2024].

[11] https://www.nomenclature-pcs.fr, [last accessed Sept. 2024].

[12] https://entreprendre.service-public.fr, [last accessed Sept. 2024].

[13] G. Mater, C. Paris, and J. Lavoué, "Descriptive analysis and comparison of two French occupational exposure databases: COLCHIC and SCOLA", *Am. J. Ind. Med.*, vol. 59, no. 5, pp. 379-391, May 2016, doi: 10.1002/ajim.22569.

[14] https://www.inrs.fr

[15] J. Févotte et al., "Matgéné: A Program to Develop Job-Exposure Matrices in the General Population in France", *Ann. Occup. Hyg.*, vol. 55, no. 8, pp. 865-878, Sept. 2011, doi: 10.1093/annhyg/mer067.

[16] https://www.service-public.fr, [last accessed Sept. 2024].

[17] https://docs.posit.co/ide/user/, [last accessed Sept. 2024].

[18] Z. Zhenhua, A. Dutta, and F. Dai, "Exoskeletons for manual material handling – A review and implication for construction applications", *Autom. Constr.*, vol. 122, Feb. 2021, doi: 10.1016/j.autcon.2020.103493.

[19] A. J. Williams, J. C. Lambert, K. Thayer, and J.-L. C. M. Dorne, "Sourcing data on chemical properties and hazard data from the US-EPA CompTox Chemicals Dashboard: A practical guide for human risk assessment", *Environ. Int.*, vol. 154, pp. 106566, Sept. 2021, doi: 10.1016/j.envint.2021.106566.

[20] C. Fourneau et al., "The French 2016-2020 National Occupational Health Plan: a better understanding of multiple exposures", *Environ. Risques Santé*, vol. 20, no. 4, pp. 377-382, Aug. 2021, doi: 10.1684/ers.2021.1570.

[21] R. R. Boyles, A. E. Thessen, A. Waldrop, and M. A. Haendel, "Ontology-based data integration for advancing toxicological knowledge", *Curr. Opin. Toxicol.*, vol. 16, pp. 67-74, Aug. 2019, doi: 10.1016/j.cotox.2019.05.005.

[22] R. R. Rao, K. Makkithaya, and N. Gupta, "Ontology based semantic representation for Public Health data integration", in *2014 International Conference on Contemporary Computing and Informatics (IC3I)*, p. 357-362, Nov. 2014, doi: 10.1109/IC3I.2014.7019