

Modélisation des données ouvertes au service des ODD : application aux événements sportifs en ville

Wissal Benjira^{*,**}, Faten Atigui^{***}, Bénédicte Bucher^{**},
Malika Grim-Yefsah^{**}, Nicolas Travers^{*,***}

* De Vinci Higher Education, DVRC Research Center, Paris, France
{prénom.nom}@devinci.fr

** LASTIG, Université Gustave Eiffel, IGN, Saint Mandé, France
{prénom.nom}@ign.fr

*** CEDRIC, Conservatoire National des Arts et Métiers, Paris, France
{prénom.nom}@cnam.fr

Résumé. Atteindre les Objectifs de Développement Durable (ODD) est un enjeu majeur à l'horizon de 2030. Pour évaluer les progrès réalisés, des indicateurs sont indispensables et doivent pouvoir être comparables aussi bien dans le temps que dans l'espace. La définition de ces indicateurs nécessite l'exploitation de données existantes notamment des données ouvertes. Toutefois, la diversité des sources et de formats pose des défis en termes de structuration et d'intégration de données. Malgré l'abondance des données ouvertes, leur utilisation reste limitée, laissant un potentiel inexploité pour orienter les politiques urbaines vers la durabilité. Ainsi, cet article présente une approche de modélisation dirigée par les données pour représenter les données ouvertes. Nous illustrons l'application de cette méthode aux ODD, et aux événements sportifs ouvrant ainsi de nouvelles perspectives pour établir le rapprochement entre données et calcul d'indicateurs.

1 Introduction

Les Objectifs de Développement Durable (ODD) établis par les Nations Unies représentent un appel global à l'action et définissent une feuille de route destinée à répondre aux défis de notre époque. À cet égard, il est essentiel de s'appuyer sur des informations précises et pertinentes pour éclairer la prise de décision. Les 17 ODD sont associés à des cibles plus précises et sont munis d'indicateurs permettant de suivre les progrès. Définis à différentes échelles, les indicateurs doivent être suffisamment comparables dans le temps et dans l'espace.

Cependant, la diversité des sources de données et des formats pose un défi majeur en matière de structuration, de normalisation et d'intégration des données. Face à cette complexité, des organisations internationales incitent les pays à renforcer la documentation et la collecte de données notamment au niveau urbain (Nations-Unies, 2015).

Abritant plus de la moitié de la population mondiale, les villes sont couvertes par de nombreuses données ouvertes notamment des données ouvertes. Ces données représentent un gisement à considérer et à exploiter pour guider les politiques urbaines vers des voies plus durables et rési-

lientes. La modélisation des données ouvertes revêt ainsi une importance capitale, en particulier dans les villes qui sont le laboratoire de ce développement, confrontées à une multitude de complications, mais aussi d'opportunités.

Parmi les nombreuses opportunités, les Événements Sportifs (ES) se démarquent comme des cas d'application emblématiques. Ces manifestations, championnats sportifs nationaux ou internationaux, créent un environnement complexe avec des impératifs de logistique, de sécurité et de durabilité. Les ES peuvent servir de catalyseur pour des projets visant la durabilité.

Par conséquent, cet article propose une méthode spécifique d'aide à la structuration de données ouvertes, guidée par des exigences d'alimentation d'indicateurs ODD. Cette approche vise à créer des schémas de données cohérents, à définir des normes et à construire des cadres pour transformer des données disparates en informations utiles.

Cet article est organisé en trois sections. Dans la section 2, nous effectuons une analyse approfondie de l'état de l'art dans le domaine de la structuration des données ouvertes en relation avec les ODD, tout en examinant les modèles associés aux ES urbains. Par la suite (section 3), nous présentons notre méthodologie de modélisation en décrivant sa mise en œuvre pratique. Enfin, en section 4 nous rapportons les conclusions et perspectives.

2 État de l'art

Les ODD offrent une grille d'analyse dans laquelle il est pertinent de projeter les prises de décisions. Cela représente un défi auquel le croisement de données peut répondre. Diverses activités de recherche ont tenté de développer et d'appliquer des techniques capables de gérer les complexités associées aux données des ODD (Howells et al., 2013; Joshi et al., 2015; Kumar et al., 2018; Almannai et al., 2020). Toutefois, très peu d'études ont proposé des modélisations informatiques pour représenter leurs liens avec ES.

2.1 Défis et opportunités dans la modélisation de données pour les ODD

Les efforts visant à atteindre les ODD reposent sur la modélisation et la simulation, créant ainsi des systèmes d'aide à la décision basés sur des modèles (Allen et al., 2016; Abson et al., 2017; Arnold et al., 2020). Pereira et al. (2021) ont introduit un cadre basé sur la théorie des graphes pour modéliser les ODD, offrant une vision systémique de leurs interdépendances.

La modélisation des données ouvertes pour les ODD englobe diverses sources (données gouvernementales ou citoyennes). Cette approche repose sur la transparence et l'accessibilité des données pour des décisions éclairées (Joshi et al., 2020). Ainsi, le web sémantique offre un cadre pour organiser et interconnecter ces données, facilitant l'analyse (Joshi et al., 2020).

En 2017, les Nations Unies ont élaboré une ontologie des données basée sur le cadre mondial des indicateurs des ODD (Joshi et al., 2021). Cette dernière vise à faciliter l'accès aux informations sur les ODD et à intégrer différentes sources dans les applications utilisateur, comprenant des identificateurs pour les objectifs, les cibles et les indicateurs de l'Agenda 2030.

En France, l'INSEE¹ coordonne la collecte d'indicateurs statistiques pour suivre les progrès. Le CNIS² a mis en place un groupe de travail dédié aux ODD, mettant en lumière des

1. INSEE - Institut national de la statistique et des études économiques

2. CNIS - Conseil National de l'Information et de la Statistique

cibles prioritaires dépourvues d'indicateurs. Des recommandations, notamment pour l'ODD 11 (qui concerne les villes et les communautés durables) guident vers les indicateurs prioritaires.

En particulier, un travail de recherche est suggéré autour d'indicateurs d'accessibilité. Nous pouvons l'estimer en tenant compte des caractéristiques urbaines qui définissent ou influencent les conditions de circulation, en particulier : la distance par rapport aux transports publics, l'employabilité et la diversité de l'occupation des sols (Hanani et Aziz, 2021). Ces trois facteurs sont recensés par les sources de données ouvertes de l'INSEE.

2.2 Approches informatiques pour la modélisation des ES en ville

Depuis la fin du XXe siècle, les petits, moyens et grands ES se sont considérablement développés. Les travaux existants dans ce domaine se concentrent sur plusieurs axes de recherche. Tout d'abord, la modélisation des flux de spectateurs est étudiée pour anticiper et gérer les déplacements des visiteurs, minimisant ainsi les congestions de trafic et les problèmes de sécurité (Smith, 2014). Les techniques de simulation, telles que les modèles basés sur des agents, sont couramment utilisées pour simuler les mouvements des foules (Karamouzas et al., 2014).

En outre, la modélisation des infrastructures sportives et des sites d'accueil joue un rôle crucial dans la planification des événements. Grim-Yefsah et Bucher (2019); Grim-Yefsah et al. (2020) ont proposé un modèle conceptuel permettant de représenter les caractéristiques d'un ES et les liens entre ce dernier et les infrastructures sportives. Wu et al. (2017); Kurowska et al. (2021) ont utilisé les systèmes d'information géographique pour cartographier ces infrastructures, en tenant compte de facteurs tels que la capacité, la localisation et l'accessibilité.

Ainsi, les enjeux clés de l'Agenda 2030 en matière de données résident dans la nécessité d'assurer la disponibilité de données de haute qualité, actualisées, désagrégées et fiables. Notre approche consiste à proposer une méthode de production d'un modèle du domaine des ODD pour guider la réutilisation et la structuration des données ouvertes, étudiant les impacts des ES sur différentes dimensions des ODD, tout en tenant compte de la diversité des sources de données et en définissant une structuration cohérente (Preuss, 2019).

3 Approche de modélisation des données

La modélisation des données ouvertes en vue de les intégrer dans un système de gestion ou d'analyse nécessite une démarche systématique et méthodique. Nous avons élaboré une démarche en quatre étapes pour guider ce processus (fig. 1). Dans cette section, nous présentons la démarche suivie en section 3.1 et les modèles conceptuels des ODD et des ES en section 3.2.

3.1 Démarche de modélisation dirigée par les données

Dans cette section, nous détaillons les étapes qui composent la démarche (fig. 1). Chaque étape est nécessaire pour aboutir à une représentation fiable à partir de données ouvertes.

1. Collecte des données : Quelles sources de données utiliser ?

L'acquisition des données implique la collecte de jeux de données et des métadonnées provenant de diverses sources, qualitatives et quantitatives en accès et usage libre. Cette étape rassemble les matériaux bruts nécessaires à la construction d'un modèle de données.

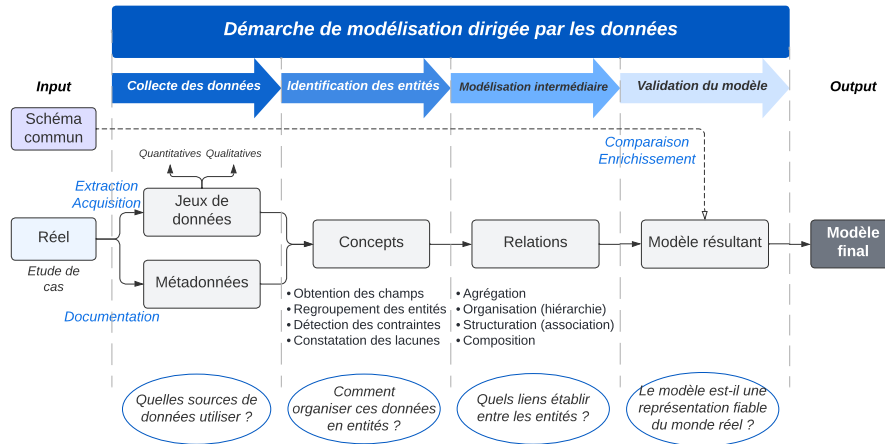


FIG. 1 – Approche dirigée par les données pour la modélisation des données ouvertes

2. Identification des entités : Comment organiser ces données en entités ?

Cette étape identifie les entités et les éléments clés dans les données. Il s'agit de sélectionner les entités pertinentes, détecter les contraintes et les lacunes en termes de qualité des données. Les métadonnées, fournissant de l'information sur la donnée, sont une source insigne.

3. Modélisation intermédiaire : Quels liens établir entre les entités ?

La structure du modèle est créée en établissant des relations, des agrégations, une structuration des associations et des liens de composition entre les entités identifiées. Cette phase nécessite une compréhension approfondie de la donnée et son contexte pour définir comment elles interagissent.

4. Validation du modèle : Le modèle est-il une représentation fiable du monde réel ?

Une comparaison est effectuée avec un schéma commun ou une ontologie existante. Cette étape vise à garantir la cohérence et la fiabilité du modèle. Le modèle est enrichi si besoin.

En suivant ces étapes, nous visons à créer des modèles de données ouvertes qui sont à la fois précis et adaptés aux besoins spécifiques. Ces étapes sont essentielles pour garantir l'intégration des données ouvertes dans le système de gestion ou d'analyse. Cette méthodologie systématique permet de convertir des données brutes en informations exploitables.

3.2 Construction de modèles conceptuels génériques

Dans cette section, nous appliquons la démarche proposée pour la construction d'un modèle conceptuel générique pour ODD et pour les ES en milieu urbain. Ces deux démarches illustrent l'importance de représenter de manière abstraite les éléments essentiels et les relations au sein de domaines variés.

1. **Collecte des données** : Pour construire le premier modèle, les données et métadonnées collectées proviennent de l'INSEE, de données gouvernementales françaises pour le suivi des ODD ainsi que des données et documentations ouvertes fournies par les Nations Unies.

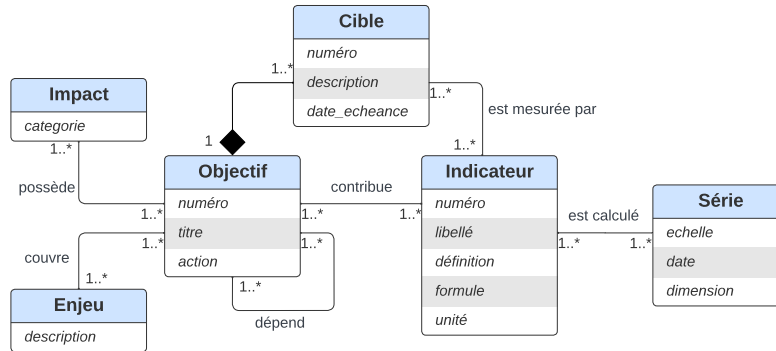


FIG. 2 – Diagramme de classe UML des ODD

2. **Identification des entités** : Les concepts clés retenus sont *impact*, *enjeu*, *objectif*, *cible*, *indicateur* et *série*.

3. **Modélisation intermédiaire** : Selon l'Agenda 2030, les *objectifs* ont un *impact* multi-dimensionnel, touchant à la fois les aspects environnementaux, économiques et sociaux. Ils couvrent un ou plusieurs *enjeux* définis par la feuille de route de l'Agenda 2030. Un *objectif* exprime un engagement ambitieux, mais spécifique, et commence toujours par un verbe ou une action. Chaque *objectif* est composé d'un certain nombre de *cibles*, qui ont des résultats quantifiables par un *indicateur* qui est une mesure précise permettant d'évaluer si un *objectif* est atteint. Chaque *objectif* est enrichi par des *indicateurs*. Dans de rares cas, un même *indicateur* peut appartenir à plusieurs *objectifs*. Enfin, une *série* est un ensemble d'observations sur une caractéristique quantitative qui fournit des mesures concrètes pour un *indicateur*. Chaque *série* contient plusieurs enregistrements de points de données organisés dans le temps, dans des zones géographiques ou dans d'autres dimensions d'intérêt (le sexe, la tranche d'âge, etc.).

4. **Validation du modèle** : Le modèle (fig. 2) a été complété de correspondances à l'aide de notions de l'ontologie *SDG Knowledge*. Le modèle conceptuel (fig. 2) représente les éléments du domaine et leurs relations, en explicitant les connaissances pertinentes de ce domaine.

Analogiquement, nous appliquons la méthode (fig. 1) pour construire un modèle conceptuel UML destiné à traiter la structuration des données relatives aux ES et à l'accessibilité.

1. **Collecte des données** : Les données et métadonnées collectées proviennent de :

- DataGouv : Recensement des équipements sportifs, espaces et sites de pratiques (2021)
- Île-de-France Mobilités : Gares et stations du réseau ferré d'Île-de-France (2023)
- IGN : Référentiel géographique du territoire concernant les plus récentes zones IRIS (2023)
- IGN : Description linéaire du réseau ferré d'Île-de-France IGN (2023)
- INJEP : Données des fédérations sportives (2022)
- INSEE : Données de la Base permanente des équipements en France (2021)
- Ministère des Sports : Base de données des familles et des types d'équipements sportifs (2023)

2. **Identification des entités** : Les concepts principaux retenus sont *fédération*, *événement sportif*, *équipement sportif*, *installation*, *adresse* et *transport*.

Modélisation des données ouvertes au service des ODD

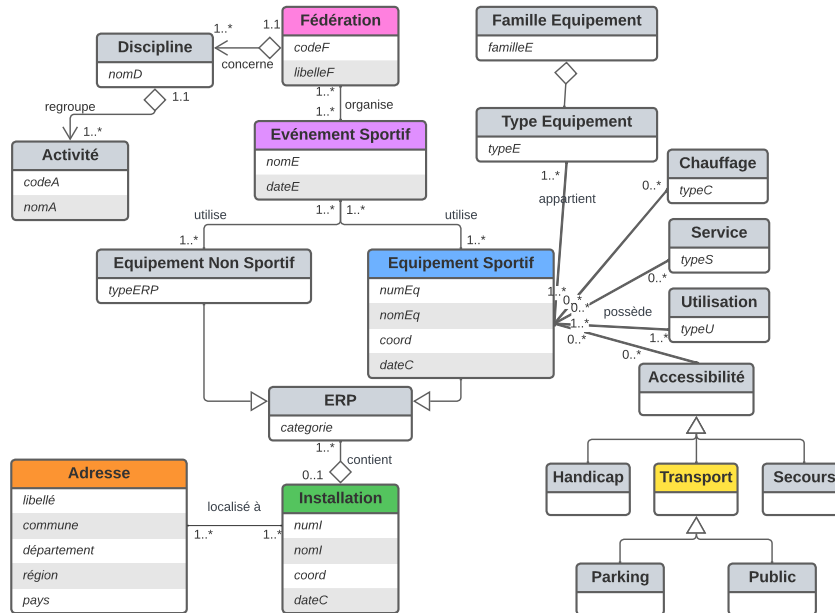


FIG. 3 – Diagramme de classes UML des ES

3. **Modélisation intermédiaire** : Le modèle intègre une multitude d’entités, de relations et d’attributs, permettant de capturer les différents aspects des ES, allant des informations géographiques aux détails des équipements sportifs, en passant par les données relatives aux fédérations sportives et aux participants. Ce modèle met en évidence l’interconnectivité de ces éléments, soulignant leur impact sur la planification, la logistique et la gestion des ES.

4. **Validation du modèle** : Le modèle (fig. 3) a été complété par le modèle de données des équipements collectifs publics provenant du référentiel de schémas de données publiques. Le modèle conceptuel présenté en figure 3 offre une vision holistique des données des ES.

L’objectif final de ces modèles est de créer un cadre permettant d’évaluer les indicateurs des ODD. Pour ce faire, la prochaine étape implique la transformation des données en un graphe. En effet, les bases de données orientées graphe se révèlent particulièrement adaptées à la gestion des relations entre les entités. Le processus de mise en correspondance des données vers le graphe n’est pas détaillé dans cet article, mais une illustration de notre approche à travers une instance est présentée (fig. 4).

Le modèle conceptuel élaboré, illustré dans la figure 3, a été appliqué pour le cas d’étude de du Championnat d’Europe de Football 2016. Le modèle a été converti en un graphe de données par un mapping “direct”. Stocké dans *Neo4j*, ce graphe a été exploré à l’aide de requêtes *Cypher*, offrant une vue globale des entités interconnectées. L’outil a été particulièrement efficace pour analyser les indicateurs ODD d’accessibilité et pour évaluer les variations géographiques. Cette approche a permis d’identifier des tendances démographiques et des zones d’impact significatif de l’événement, offrant ainsi une perspective éclairante en lien avec les cibles ODD.

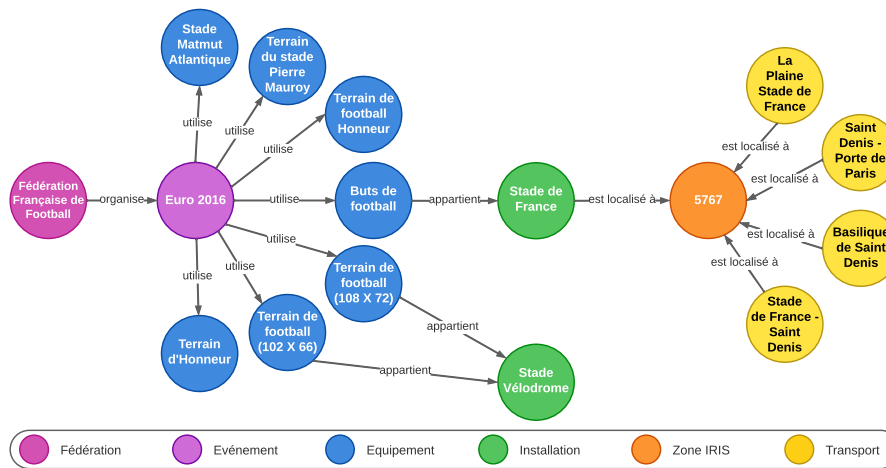


FIG. 4 – Sous-graphe Neo4j autour de l'ES du Championnat d'Europe de Football 2016

4 Conclusion et perspectives

Dans cet article, nous avons présenté une approche de modélisation dirigée par les données destinée à répondre aux besoins du développement durable. Nous avons proposé un modèle conceptuel générique capable de représenter de manière structurée une diversité de données. Nous avons illustré l'application de notre modèle à travers l'exemple du Championnat d'Europe de Football 2016. Nos futures recherches se concentrent sur le rapprochement entre ODD et ES, visant à établir des règles d'inférences sur le graphe actuel et les indicateurs de développement durable, ainsi que sur la manipulation du graphe pour produire des indicateurs.

Références

- Abson, D. J., J. Fischer, J. Leventon, J. Newig, T. Schomerus, U. Vilsmaier, H. Von Wehrden, P. Abernethy, C. D. Ives, N. W. Jager, et al. (2017). Leverage points for sustainability transformation. *Ambio* 46, 30–39.
- Allen, C., G. Metternicht, et T. Wiedmann (2016). National pathways to the sdgs : A comparative review of scenario modelling tools. *Environmental Science & Policy* 66, 199–207.
- Almannaei, N. A., M. S. Akhter, et A. Shah (2020). Improving environmental policy-making process to enable achievement of sdgs. *EPL* 50(1-2), 47–54.
- Arnold, T., J. H. Guillaume, T. J. Lahtinen, et R. W. Vervoort (2020). From ad-hoc modelling to strategic infrastructure : A manifesto for model management. *iEMSs* 123, 104563.
- Grim-Yefsah, M. et B. Bucher (2019). Towards improving knowledge capitalization system for sport events legacy. In *KMIS @ IC3K'19*, Vienna, France.

- Grim-Yefsah, M., B. Jadi, P. Thibault, et V. Arthur (2020). Modeling sport events legacy based on olympic games. In *KMIS @ IC3K'20*.
- Hanani, F. et S. Aziz (2021). Improving traffic congestion assessment by using fuzzy logic approach. *JATIT Vol.99. No 3*, 625–638.
- Howells, M., S. Hermann, M. Welsch, M. Bazilian, R. Segerström, T. Alfstad, D. Gielen, H. Rogner, G. Fischer, H. Van Velthuisen, et al. (2013). Integrated analysis of climate change, land-use, energy and water strategies. *Nature Climate Change* 3(7), 621–626.
- Joshi, A., S. Lovell, L. G. G. Morales, et A. Helton (2020). Connecting SDGs-related Information Resources with Linked Data. Technical report, United Nations.
- Joshi, A., L. G. Morales, S. Klarman, A. Stellato, A. Helton, S. Lovell, et A. Haczek (2021). A knowledge organization system for the un sdgs. In *The Semantic Web*, pp. 548–564.
- Joshi, D. K., B. B. Hughes, et T. D. Sisk (2015). Improving governance for the post-2015 sdgs : scenario forecasting the next 50 years. *World Development* 70, 286–302.
- Karamouzas, I., B. Skinner, et S. J. Guy (2014). Universal power law governing pedestrian interactions. *Physical Review Letters* 113(23), 238701.
- Kumar, P., F. Ahmed, R. K. Singh, et P. Sinha (2018). Determination of hierarchical relationships among sdgs using interpretive structural modeling. *J. EDS* 20, 2119–2137.
- Kurowska, K., R. Marks-Bielska, S. Bielski, A. Aleknavičius, et C. Kowalczyk (2021). Geographic information systems and the sustainable development of rural areas. *Land* 10.
- Nations-Unies (2015). Transforming our world : the 2030 agenda for sustainable development.
- Pereira, G., A. González, et G. Blanco (2021). Complexity Measures for the Analysis of SDG Interlinkages : A Methodological Approach. pp. 13–24.
- Preuss, H. (2019). Event legacy framework and measurement, international journal sport policy & politics. *Int. J. Sport Policy Politics* 11(1), 103–118.
- Smith, A. (2014). Leveraging sport mega-events : new model or convenient justification? *Journal of Policy Research in Tourism, Leisure and Events* 6(1), 15–30.
- Wu, X., S. V. Ukkusuri, et Y. Son (2017). Assessing Olympic transportation legacy : A comparison of performance metrics across different games. *J. Transport Geography* 59, 99–111.

Summary

Achieving the Sustainable Development Goals (SDGs) is a major challenge that must be met by 2030. To succeed, many indicators are needed, and these must be compared both in time and space, in order to assess progress. The definition of these indicators requires the use of existing data, in particular open data. However, the diversity of data sources and formats raises major challenges in terms of structuring and integration. Despite the abundance of open data, its exploitation remains limited, leaving untapped potential for guiding urban policies towards sustainability. This article presents a data-driven modeling approach to represent such data and to map it into graph database. We illustrate the application of this method to the SDGs, and to sporting events, opening up new perspectives for bridging the gap between data and indicator calculation.