Sémantisation du contenu et du contexte pour la gestion de l'expérience utilisateur et de la multimodalité en environnement de Réalité Virtuelle et Augmentée

Nicolas Saint-Léger | *Nicolas Férey* | *Joe Raad* | *Patrick Bourdot*

nicolas.saint-leger@universite-paris-saclay.fr | *nicolas.ferey@universite-paris-saclay.fr* | *joe.raad@universite-paris-saclay.fr* | *patrick.bourdot@cnrs.fr*

Motivation

L'utilisation de la sémantique dans les environnements immersifs tels que la Réalité Virtuelle (RV) et la Réalité Augmentée (RA) offre de nouvelles opportunités pour améliorer l'interaction utilisateur au travers d'interfaces multimodales, comprenant les gestes et les commandes vocales. Cependant, des défis importants subsistent en raison du **manque de représentations sémantiques formelles et explicites des contextes d'interaction et du contenu de la scène virtuelle**.

L'interaction multimodale canonique du « Put-That-There », énoncée par Bolt en 1980 [1], n'a finalement pas trouvé jusqu'à ce jour une solution générique satisfaisante, en particulier dans des situations immersives. En l'occurrence, **les approches antérieures sont généralement peu paramétrables, avec un vocabulaire de paroles et de gestes limités** et appliquées à des scénarios très définis.

Néanmoins**, l'utilisation de graphes de connaissances dans un contexte immersif permettrait de lier le contenu d'une scène virtuelle avec le contexte d'interaction**, offrant la possibilité d'améliorer l'adaptabilité des interactions multimodales, et pouvant ainsi être appliqué à n'importe quel domaine métier, tout en restant réalisable en temps interactifs. Ce travail présente une approche visant à transformer une simulation virtuelle en graphe de connaissance, permettant des questionnements et des analyses complexes de la scène virtuelle, notamment pour la génération de commande multimodale.

État de l'Art sur l'utilisation de la sémantique en environnements immersifs

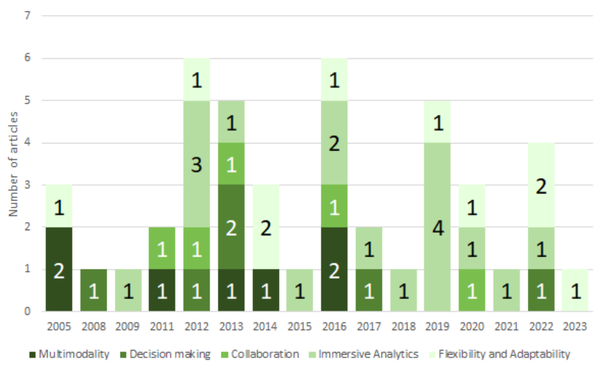

Figure 1 : Nombre d'articles classifié par motivation et groupé par date

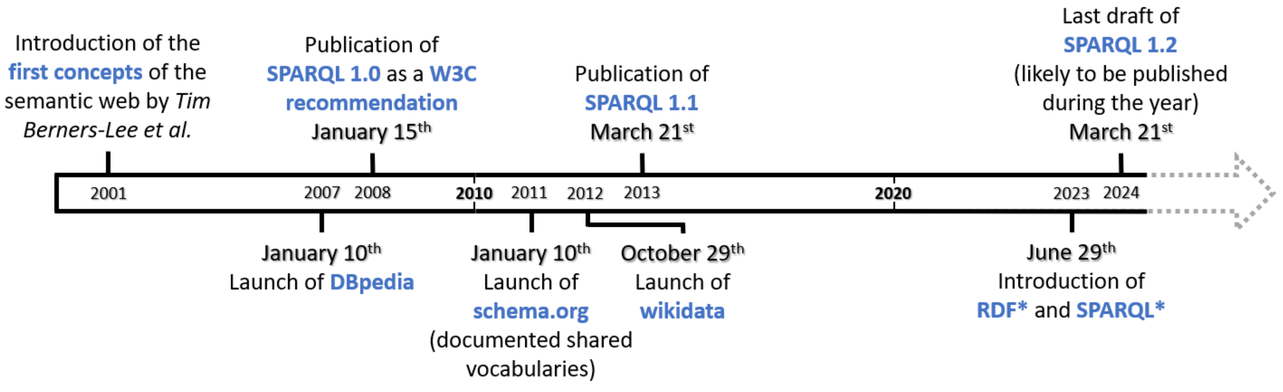

Figure 2 : Évolution du web sémantique depuis son introduction en 2001 par Berners-Lee et al. [2].

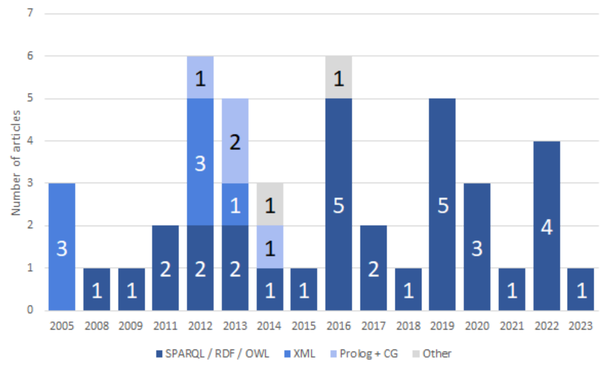

Figure 3 : Nombre d'articles classifié par technologie sémantique et groupé par date

- 5 motivations de sémantiser des scènes
  - ➢ Multimodalité
  - ➢ Prise de décision
  - ➢ Collaboration
  - ➢ Analyse immersive
  - ➢ Flexibilité et adaptation

Dans la **figure 1**, on constate qu'**il n'y a plus de travaux majeurs utilisant la sémantique pour faire de la multimodalité en immersion à partir de 2016**. Alors que, comme le montre la **figure 2**, **les technologies sémantiques continuent d'évoluer** notamment avec la standardisation des graphes de connaissance grâce au **SPARQL 1.1 (en 2013)** et plus récemment **SPARQL 1.2 (en 2024)**, comme nous le confirme la **figure 3**.

Les technologies sémantiques se sont standardisées au fil du temps, aboutissant à l'adoption de SPARQL / RDF / OWL, tirant parti du **concept de monde ouvert**, permettant une **interprétation flexible et évolutive des données**.

Architecture & Multimodalité

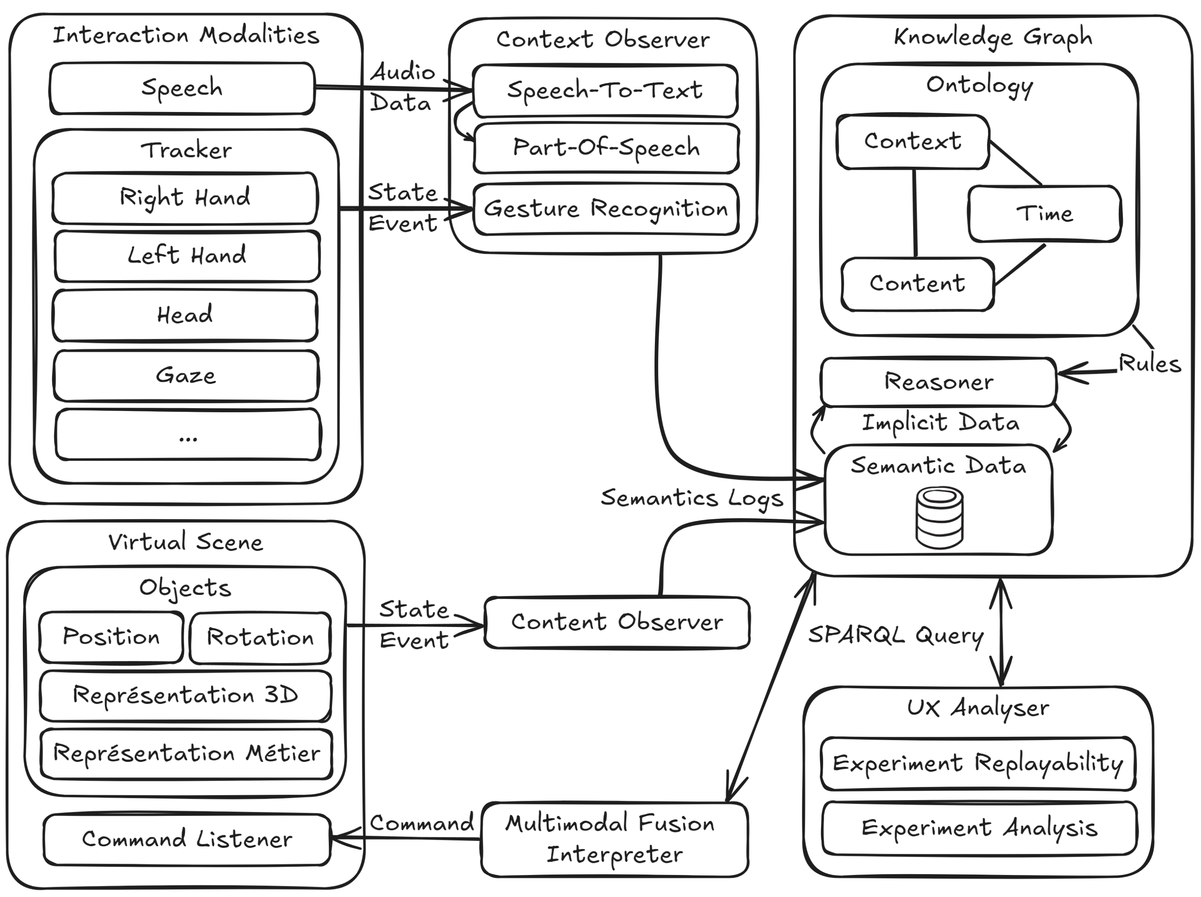

Figure 4 : Architecture de sémantisation du contenu de scène virtuel et du contexte d'interaction pour la génération de commande multimodale et la rejouabilité d'expérience

La multimodalité englobe l'utilisation de plusieurs canaux sensorimoteurs pour traiter les commandes des utilisateurs et restituer des retours visuels, auditifs ou haptiques.

- **Sortie multimodale** : Le système fournit des résultats sous différents formats sensoriels. Par exemple, Ferey et al. [3] ont développé une interaction en VR pour l'amarrage de protéines, utilisant un **rendu visuel** pour représenter les protéines, un **retour haptique** pour simuler les collisions, et un **retour audio** pour des indices supplémentaires.
- **Entrée multimodale** : L'interaction multimodale consiste à **analyser et interpréter plusieurs canaux de communication** en temps réel [4]. Le défi est de fusionner ces événements en commandes interprétables par l'application. Oviatt et al. [5] identifient deux types d'architectures pour cette fusion : celle intégrant les signaux au niveau des caractéristiques (« early fusion ») et celle les intégrant au niveau sémantique (« late fusion »).

Sémantique & Ontologie

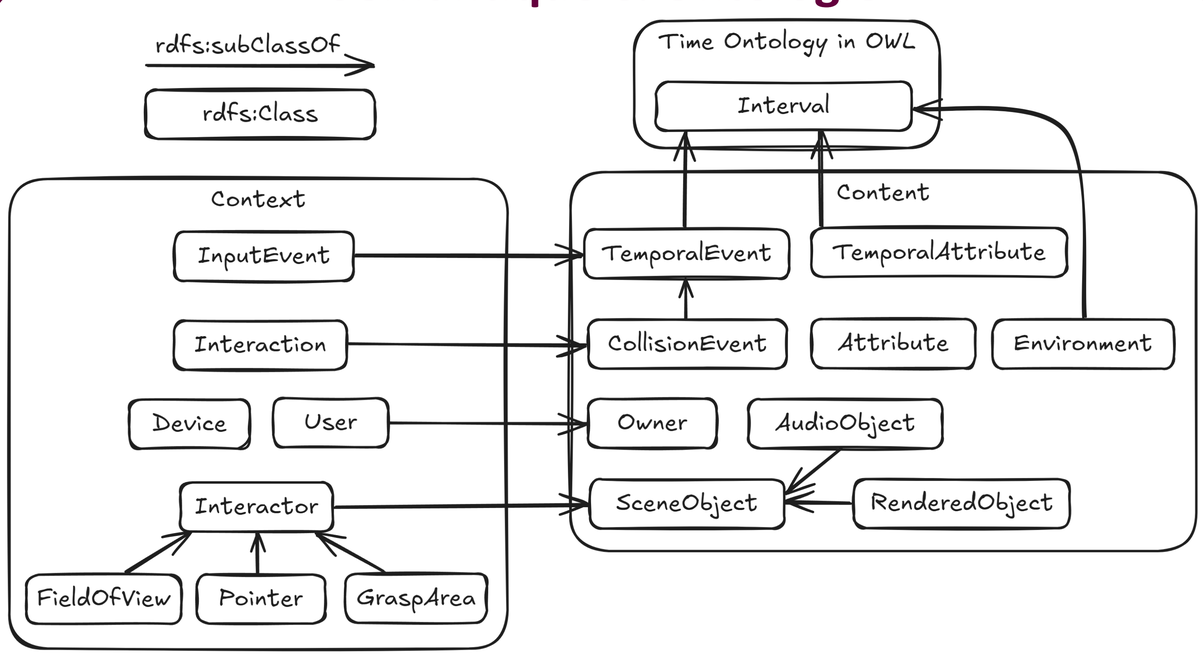

Figure 5 : Ontologie modélisant au fil du temps le contenu de scène virtuelle, ainsi que le contexte d'interaction

- **Time Ontology in OWL** [6] : Ontologie standardisée pour la modélisation temporel, grâce à des concepts tels que les **instants, intervalles et durées**.
- **Contenu** : Événements, représentations et conformations des objets d'une scène virtuelle à chaque instant donnée, grâce à des **concepts génériques** tels que les positions, tailles et orientations, ou bien des **concepts de domaine métiers**.
- **Contexte** : Interactions réalisées par les utilisateurs sur le contenu de scène, grâce à des concepts tels que le **pointage, la vision, la préhension ou la parole**.

Perspectives

- Permettre la transformation de n'importe quelle expérience virtuelle en graphe de connaissance, tout d'abord de manière **générique** avec des objets prototypiques, puis avec des domaines métiers tels que la **biologie moléculaire** [7], ou la **météorologie** [8].
- Rendre les **scènes virtuelles rejouables et analysables** à postériori d'expérience.
- Permettre la **génération de commande multimodale en temps réel**, combinant les gestes et la parole.
  - ➢ Comment interpréter la parole ?
  - ➢ Comment capter l'intention de l'utilisateur ?
  - ➢ Comment gérer la multimodalité en collaboration ?

Références

1. R. A. Bolt (1980, July). "Put-that-there" Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques* (pp. 262-270).
2. T. Berners-Lee, J. Hendler and O. Lassila (2001). A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific american*, *284*(5), 34-43.
3. N. Férey, J. Nelson, C. Martin, L. Picinali, G. Bouyer, A. Tek, P. Bourdot, J.-M. Burkhardt, B. F. Katz, M. Ammi, et al. (2009). Multisensory VR interaction for protein-docking in the CoRSAIRe project. *Virtual Reality*, *13*, 273-293.
4. P. Martin, A. Tseu, N. Férey, D. Touraine and P. Bourdot (2014, February). A hardware and software architecture to deal with multimodal and collaborative interactions in multiuser virtual reality environments. In *The Engineering Reality of Virtual Reality 2014* (Vol. 9012, pp. 68-83). SPIE.
5. S. Oviatt, P. Cohen, L. Wu, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson, et al (2000). Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions. *Human-computer interaction*, *15*(4), 263-322.
6. F. Pan and J. R. Hobbs (2006). Time Ontology in OWL. *W3C working draft, W3C*, *1*(1), 1.
7. M. Trellet, N. Férey, J. Flotyński, M. Baaden and P. Bourdot (2018). Semantics for an integrative and immersive pipeline combining visualization and analysis of molecular data. *Journal of integrative bioinformatics*, *15*(2), 20180004.
8. I. Ouedraogo, H. Nguyen and P. Bourdot (2024). Immersive analytics with augmented reality in meteorology: an exploratory study on ontology and linked data. *Virtual Reality*, *28*(3), 144.