Graphes de connaissances pour les humanités numériques : besoins spécifiques et problèmes généraux

Knowledge graphs for the digital humanities: specific requirements and general issues

Mathieu d'Aquin[1]

[1]*Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France*

Abstract

A lot has been written about the benefits and difficulties associated with the creation and use of knowledge graphs and of the broader Semantic Web technologies in the context of the digital humanities. Concrete feedback from projects using such technologies is however rare and, by nature, often focuses on the specific use of knowledge graphs in a particular context. In this article, we describe the use of Semantic Web technologies in three different domains, from three (multidisciplinary) projects. The goal is to better understand how, despite the variety of needs and requirements from researchers in those projects, shared benefits and issues appear. Identifying those shared elements can be helpful in the sense that it can guide the development of Semantic Web tools towards greater relevance and efficacy for the digital humanities.

Keywords

knowledge graphs, Semantic Web, music history, iconography, literature

1. Introduction

Le Web sémantique peut être vu comme une utilisation des technologies du Web pour rendre les données et les connaissances plus accessibles et manipulables au travers de graphes de connaissances (voir [1]) inter-connectées et navigables par des requêtes HTTP. S'ajoute à cela la spécification de la signification de ces données et de ces connaissances par l'utilisation d'ontologies (voir par exemple [2]). Ce type de représentation permet d'encoder les informations relatives à un domaine non seulement de façon à être connecté avec d'autres sources d'information, mais aussi au travers de représentations flexibles et évolutives.

Du fait de leur nature pluridisciplinaire et des sujets complexes qui y sont étudiés, les informations manipulées par les disciplines des humanités numériques sont souvent très riches et variables. Elles incluent un grand nombre de dimensions, instanciées de façon plus ou moins complète, et peuvent faire référence à des sources externes, telles que des bases de données de référence dans le domaine. Pour cette raison, les humanités numériques ont été depuis longtemps considérées comme un champ d'application privilégié pour les technologies du Web sémantique (voir par exemple [3]).

Beaucoup a été dit sur les bénéfices et les difficultés rencontrés dans la mise en place de ces technologies dans des applications des humanités numériques. De nombreux cas d'utilisation

. *Workshop on Digital Humanities and Semantic Web*
. mathieu.daquin@loria.fr (M. d'Aquin)
. https://mdaquin.github.io/ (M. d'Aquin)

ont été décrits, mais les retours d'expérience concrets restent rares [1]. De plus, à part quelques exceptions où les sujets généraux de l'utilisation des technologies du Web sémantique ont été explicitement étudiés comme dans ([4]), ces retours par nature ont tendance à se focaliser sur une utilisation spécifique des graphes de connaissances dans un domaine particulier.

Cet article décrit l'utilisation des technologies du Web Sémantique dans trois domaines différents, issus de trois projets pluridisciplinaires différents :

L'iconographie avec un projet de portail fondé sur un graphe de connaissances décrivant les fresques murales d'églises de Crète.

L'histoire de la musique avec un projet d'acquisition d'une base de milliers de descriptions d'expériences d'écoute de la musique.

La littérature avec l'analyse sémantique des écris de certaines des premières femmes philosophes.

L'objectif est de comprendre comment, malgré des besoins et des attentes variés des chercheurs de ces domaines, émergent aussi bien des avantages que des problèmes communs. Le but est donc de discuter, d'une façon indépendante du domaine d'application spécifique, quels sont les avantages réels tirés des technologies du Web sémantique, et comment celles-ci doivent encore évoluer pour devenir de meilleurs outils pour les humanités numériques.

2. Graphes de connaissances et humanités numériques : avantages attendus et blocages

De façon générale, l'idée des graphes de connaissances (voir [1]) est de représenter les données, informations et connaissances sous la forme de graphes orientés et étiquetés. Les graphes de connaissances suivent la vision générale du Web sémantique en s'appuyant sur ces technologies (HTTP, RDF [2]) pour représenter ces graphes de façon à les rendre compatibles avec une diffusion sur le Web, et avec la possibilité de les interconnecter globalement avec d'autres graphes. Ceux-ci sont aussi qualifiés de graphes de connaissances du fait que la signification des éléments d'information inclus peut être rendue explicite et interrogeable au travers d'ontologies (voir [5]), permettant ainsi de faciliter l'échange de connaissances et le raisonnement sur ces connaissances.

Comme décrit plus haut, de par leur flexibilité et leur ouverture, les approches fondées sur les graphes de connaissances ont gagné en popularité dans plusieurs domaines des humanités numériques depuis de nombreuses années. Dans ces domaines (voir par exemple [3]) et dans des domaines connexes (voir par exemple [6]), les avantages attendus souvent cités inclus :

Applications intelligentes : l'idée de représenter les connaissances au travers d'ontologies provient du domaine des systèmes à base de connaissances, c'est-à-dire d'un certain paradigme de l'intelligence artificielle s'appuyant sur la représentation explicite de connaissances et sur le raisonnement artificiel pour faciliter la prise de décision. Un des avantages souvent cité du Web sémantique est donc de permettre la construction, en suivant ce

1. Voir par exemple le workshop WHISE http://whise.cc/
2. https://www.w3.org/TR/rdf11-concepts/

paradigme, d'applications intelligentes qui exploitent les connaissances incluses dans les graphes de connaissances construits et dans ceux auxquels on se connecte (voir par exemple [7]).

Accès à l'information : à un niveau plus bas, le simple fait de représenter les informations et les connaissances du domaine d'une façon compatible avec les technologies du Web apparaît comme un avantage. Ces technologies sont ouvertes et conçues pour faciliter l'accès de n'importe en ligne, sans contrainte logicielle, ce qui les rend attractives pour des activités de recherche utilisant ces connaissances en comparaison avec l'utilisation de systèmes plus fermés tels que des bases de données relationnelles.

Données liées et intéroperabilité : une des notions centrales du Web sémantique est de permettre de connecter ses propres données avec d'autres, de la même façon que l'on connecte des pages Web entre elles. Cela semble particulièrement utile dans le contexte des humanités numériques considérant que des données de référence existent sous forme de graphes de connaissances (voir par exemple [8]) et peuvent donc être réutilisées. Ces graphes de connaissances utilisant non seulement des technologies standards, mais aussi des vocabulaires et ontologies partagées, il devient possible de créer de nouveaux graphes de connaissances syntaxiquement et sémantiquement interopérables : ils peuvent être utilisés conjointement dans des applications sans difficultés majeures.

Visibilité : beaucoup de projets d'humanités numériques utilisant les graphes de connaissances sont soit liés au patrimoine culturel, soit liés à des activités de recherche dans le domaine concerné. Dans les deux cas, en plus de l'avantage de pouvoir se connecter à des sources d'informations externes comme décrit ci-dessus, l'utilisation des technologies du Web sémantique a aussi généralement pour but de permettre la création de nouvelles ressources de référence, réutilisables par d'autres, et permettant ainsi une plus grande visibilité pour le projet.

Malgré ces avantages attendus, l'utilisation des technologies liées aux graphes de connaissances reste limitée au sein des humanités numériques, et cette utilisation n'est pas toujours soutenue ou pérenne. Dans sa thèse soutenue récemment, [4] s'est intéressée aux raisons pour lesquelles les chercheurs de certains domaines des humanités numériques utilisent ou n'utilisent pas les technologies du Web sémantique et quelles sont les directions possibles pour les améliorer. Parmi les blocages identifiés sont inclus :

Des technologies compliquées et mal connues : un des désavantages les plus communément cités des graphes de connaissances est qu'ils font appel à des technologies qui restent difficiles à comprendre. En effet, celles-ci sont généralement plus récentes que les alternatives possibles, moins bien documentées, et utilisables au travers d'outils souvent développés au sein d'équipes de recherche ne disposant pas de service d'aide suffisant pour permettre aux non-experts de les utiliser facilement. S'investir dans l'utilisation de ces technologies représente donc un risque qu'il est parfois difficile de prendre.

Alternatives plus accessibles : en contrepartie de ce qui est décrit ci-dessus, il est souvent possible de réaliser un projet en utilisant des technologies plus traditionnelles, telles que les systèmes de gestion de bases de données relationnelles, pour lesquels une expérience, de l'aide et un certain niveau de support sont disponibles. Les avantages cités plus haut ne

seront de fait pas réalisés, mais leur importance comparé aux avantages de l'utilisation d'outils établis reste difficile à communiquer aux collaborateurs s'intéressant aux aspects techniques du projet (comme par exemple le service informatique d'un département d'une université).

Coût : beaucoup des outils utilisés pour construire des graphes de connaissances ou pour utiliser les graphes de connaissances dans des applications sont libres et gratuits. Le coût devrait donc être considéré comme un avantage. Il faut néanmoins, comme cité ci-dessus, comparer ce coût à l'utilisation d'outils déjà établis est disponible dans l'équipe du projet. S'ajoute à cela non seulement le coût de déploiement (serveur, maintenance) mais aussi le coût en ressources humaines : il est souvent en effet nécessaire d'acquérir les compétences nécessaires à l'utilisation de ces technologies.

Le but ici est de confronter cette perception à l'expérience concrète de plusieurs projets dans des domaines variés. Les sections suivantes présentent les trois projets, et sont suivies d'une discussion sur les avantages réellement réalisés et les difficultés rencontrées dans ces trois projets, pour conclure sur le besoin d'évoluer les technologies pour lever certaines de ces difficultés et mieux mettre en avant les avantages.

3. Le projet LEDA : l'enfer et les pécheurs en Crète

Le projet LEDA [3] s'intéresse aux représentations de l'enfer dans les églises de Crète. L'intérieur de ces églises est traditionnellement recouvert de fresques murales représentant différents aspects de la religion : des personnages, des scènes bibliques, etc. La localisation de ces fresques au sein de l'église est significative, et le projet s'intéresse tout particulièrement à celles représentant l'enfer, les pécheurs et les tortures qui leur sont infligées. Un des objectifs du projet est de répertorier ces représentations pour comprendre quelles sont les conventions de représentation de l'enfer, de scènes spécifiques, des péchés, en fonction de l'église, de la région, etc. (voir [9], l'ouvrage en deux volumes publié à l'issue du projet).

L'utilisation des graphes de connaissances est ici liée à un besoin de représentation évolutive des informations et à la facilitation de la navigation dans ces informations. En effet, les chercheurs impliqués dans le projet souhaitaient avoir un portail, accessible à l'équipe de recherche et à d'autres, de façon à pouvoir explorer les milliers de photographies collectées et facilement obtenir, par exemple, toutes les représentations de la même scène ou du même péché. Il devait être possible de filtrer les représentation par église, région, localisation au sein de l'église, péché ou scène et la représentation de la localisation devait être précise : il devait clairement apparaître la partie de l'église dans laquelle se trouvait la fresque, sur quel élément architectural, à quelle hauteur, à côté de quelles autres fresques, etc. D'autres informations, telles que les fresques que représente chaque photographie, devaient aussi être présentes.

Une des raisons pour lesquelles ce projet s'est tourné vers les graphes de connaissances est lié à l'accessibilité, la flexibilité et l'évolutivité de ce mode de représentation. En effet, l'annotation et la classification des fresques et de photographies étaient préalablement réalisées au travers d'une base de données Microsoft Access (technologie connue et maîtrisée par certains membres

3. https://ledaproject.org.uk

de l'équipe). Il est apparu clairement néanmoins que la mise à disposition sur le Web d'une telle base de données serait difficile. D'une façon plus importante, un des problèmes de ce mode de représentation est qu'il était nécessaire de fixer le modèle des données à priori, alors que l'équipe n'avait pas encore finalisé les types d'explorations et d'interrogations qu'ils seraient amenés à faire au cours de leur recherche. Finalement, même si les données elles mêmes n'étaient pas de très grande taille, leur richesse (les caractéristiques potentiellement renseignées pour chaque fresque sont nombreuses) et leur hétérogénéité (beaucoup de données incomplètes, avec des valeurs variées, dans plusieurs langues) rendait la création d'une base de données "classique" et sa mise à disposition difficile.

Afin d'utiliser les avantages des technologies du Web sémantique pour pallier ces problèmes, un processus collaboratif et incrémental a été mis en place, où la création d'une ontologie pour la représentation des fresques et des objets liés, la création du portail pour explorer les données et la réflexion sur les interrogations possibles étaient considérées en parallèle et influaient les uns sur les autres. Une des difficultés liées au projet est que les chercheurs en iconographie avaient peu de connaissances en technologie, et ne s'y intéressaient pas *a priori*. Le projet s'appuyait donc sur un dialogue entre ces chercheurs et les informaticiens où différentes itérations de l'ontologie et du portail étaient discutées pour progressivement arriver à un consensus sur les fonctions et le mode de représentation requis. Un aspect intéressant de ce projet est aussi qu'une méthode d'entrée de données utilisant des tableurs, ensuite automatiquement traduits sous la forme de graphes de connaissances, a été mise en place, permettant aux chercheurs de contrôler le contenu du graphe de connaissances sans avoir à s'investir dans l'utilisation des technologies et outils associés. Le résultat de ce processus, le portail d'exploration des informations iconographiques sur l'enfer dans 92 églises de Crète, est représenté figure 1.

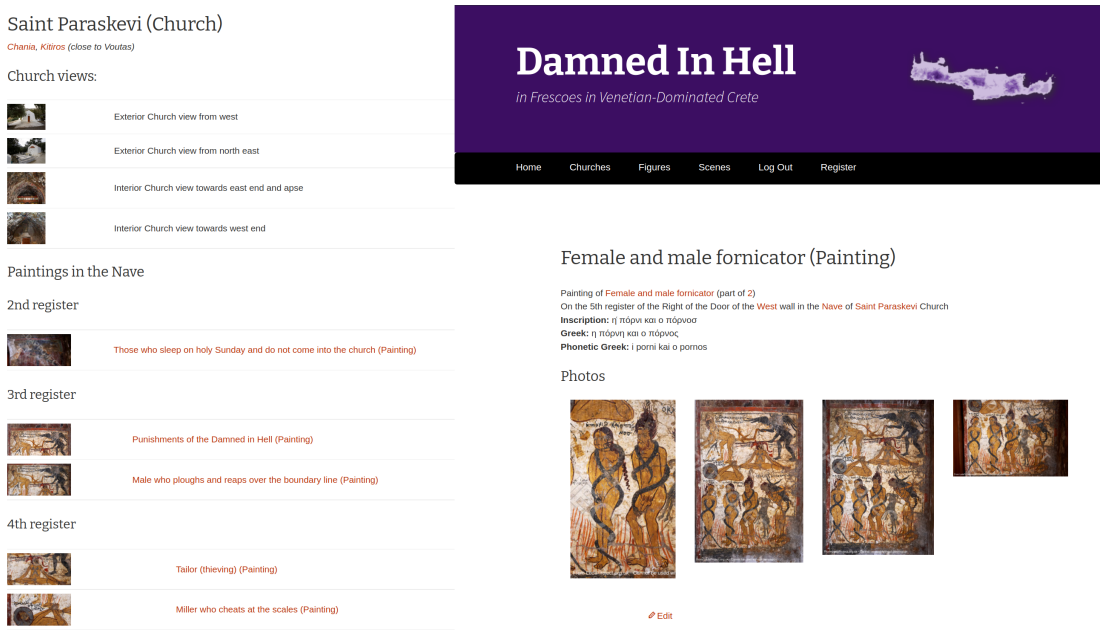

Figure 1 : Le portail du projet LEDA. Vue d'une église (à gauche) et d'une fresque (à droite).

4. Le projet *Listening Experience Database*

The Listening Experience Database (LED) [4] est un projet ambitieux dont le but est de collecter des milliers d'expériences d'écoute de la musique avec autant d'information sur les personnes impliquées, la musique, son contexte et les sources des descriptions que possible (voir [10, 11, 12]). Comme le projet précédent, une des raisons de l'utilisation des graphes de connaissances dans ce projet est la flexibilité et l'évolutivité de la représentation. En effet, les informations à collecter sont très riches, incluant des éléments spatio-temporels (le lieu et le temps de l'écoute de la musique, mais aussi de son exécution), bibliographiques, musicales, socio-économiques, etc.

De ce fait, de façon similaire au projet ci-dessus, les processus de développement de l'ontologie du projet et des outils de navigation dans les graphes de connaissances créés se sont déroulés incrémentalement, au travers d'un dialogue entre les développeurs et les équipes de recherche. De plus, on retrouve ici l'avantage associé aux données liées puisque les graphes créés se connectent à des sources d'information de référence telles que la *British Library* ou VIAF [5].

Deux autres points importants, que l'on peut voir comme des éléments de difficulté, étaient aussi à prendre en compte dans la réalisation de ce projet sur la base de graphes de connaissances : 1- certain éléments d'information était vagues, et 2- les descriptions d'expériences d'écoute de la musique devait être renseignées par les utilisateurs (variés) de la plate-forme (*crowdsourcing*). En effet, les descriptions d'expériences d'écoute (généralement un paragraphe de texte) proviennent de sources variées, incluant des correspondances, journaux personnels, etc. Par nature, ces sources ne contiennent pas toujours toutes les informations requises, mais peuvent aussi contenir des informations imprécises concernant par exemple le moment de l'écoute (*un mardi après-midi, en automne, dans les années 1920*) ou son lieu (*dans le train entre Paris et Lyon*). Pour permettre ce genre de représentations, en évitant au maximum de perdre de l'information, il a donc été nécessaire de construire des structures ontologiques riches autour de notions simples telles que le temps ou la localisation.

L'autre élément, que les données incluses devait provenir d'utilisateurs variés, a posé un certain nombre de difficultés. Encore plus que dans le projet précédent, tout d'abord, cela supposait de permettre l'édition de données d'une façon qui masque la technologie sous-jacente. Cela a été réalisé au travers d'un ensemble de formulaires dont beaucoup de champs sont optionnels et où l'entrée d'information est aussi guidée que possible. Ces formulaires permettent en particulier de réutiliser des éléments entrés par d'autres, réduisant ainsi l'effort requis et évitant les incohérences entre les entrées de différents utilisateurs.

Cet aspect de gestion des incohérences (et des erreurs) représente aussi un élément majeur de ce projet. Plusieurs utilisateurs peuvent renseigner des informations sur les mêmes personnes, musiques, livres, etc. Il se peut que des erreurs se glissent dans les contributions ou même que différents contributeurs aient différentes opinions. Pour permettre une gestion efficace de ces difficultés, un mécanisme de validation a été mis en place où chaque contributeur possède son propre graphe de connaissances, et seulement les éléments de ce graphe de connaissances qui ont été approuvés par un membre du projet sont inclus dans le graphe de connaissances général et publique du projet.

4. https://www.listeningexperience.org/
5. *The virtual authority file* - http://viaf.org/

Le résultat de ces développements est un portail (voir figure 2) qui inclut à l'heure actuelle près de 12 000 descriptions d'expériences d'écoute et qui permet de rechercher, naviguer et explorer ces descriptions en fonctions de nombreux critères (en plus d'ajouter ses propres descriptions). Ce portail permet aux chercheurs de se focaliser par exemple sur certaines périodes, certains lieux, certains genres de musique, ou certains contextes d'écoute, et d'obtenir des informations riches sur les expériences répondant à ces critères. En fournissant ainsi une plate-forme pour l'enregistrement et l'exploration de ces expériences d'écoute, LED offre ainsi un support de recherche pour de nombreux chercheurs, comme en témoignent les actes édités en ligne des deux conférences déjà organisées sur le sujet [6]. La création de cette base de descriptions d'expériences d'écoute de la musique a en plus permis le développement de nouvelles applications intelligentes, telles que *FindLEr* [7] qui permet de retrouver dans des textes quelconques des passages qui ressemblent à des descriptions d'expériences d'écoute de la musique.

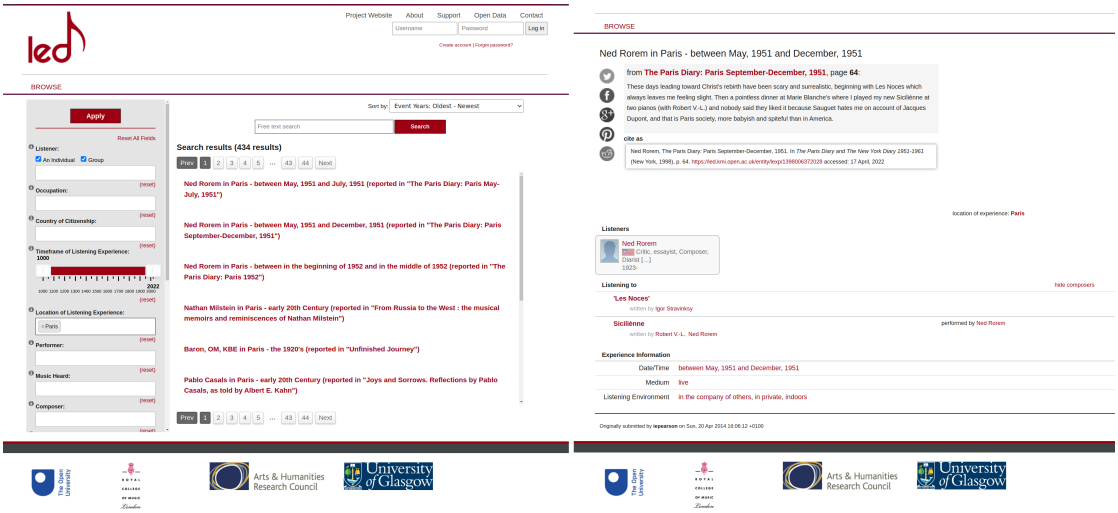

Figure 2 : Le portail du projet LED. Interface de recherche d'expériences d'écoute de la musique (à gauche - recherche d'expériences localisées à Paris) et description d'une expérience d'écoute (à droite).

5. Correspondances entre John Norris, Mary Astell et Damaris Masham

L'étude de textes philosophiques est une activité par nature complexe et sujette à interprétation. C'est d'autant plus le cas quand ces textes forment des correspondances entre plusieurs auteurs. Le but de l'outil ArguNest [8], dévelopé dans le cadre de la thèse de Ioanna Kyvernitou (NUI Galway, Irlande) est de permettre de s'abstraire des textes eux-même et de représenter, au travers de graphes de connaissances, la *lecture* de ces textes par le réseau d'arguments et de propositions qu'ils contiennent.

6. http://ledbooks.org/
7. https://led.kmi.open.ac.uk/discovery/findler
8. https://github.com/mdaquin/ArguNest

Le cœur de ce projet est donc une ontologie permettant de représenter ce réseau d'arguments et de propositions. L'analyse d'arguments est une discipline établie et de nombreux modèles de représentation de réseaux d'arguments existent. On s'intéresse ici en particulier aux lettres échangées par John Norris, un théologue reconnu de son époque, et Mary Astell et Damaris Masham, qui remettent en cause certains des raisonnements exposés dans ses ouvrages. Un aspect important dans la représentation ontologique de ces échanges était donc la capacité de référencer des arguments déjà établis, et donc de séparer la notion abstraite d'argument de sa matérialisation dans les textes. En effet, le même argument peut être réutilisé, re-discuté ou mentionné plusieurs fois dans les textes, sous plusieurs formes différentes. L'objectif était donc de pouvoir représenter une lecture particulière des textes à deux niveaux d'abstraction : les annotations des textes comme représentant des arguments et des propositions, et comment ces arguments et ces propositions, en tant qu'entités abstraites, sont liés entre eux.

L'outil créé sur la base de cette ontologie (ArguNest, voir figure 3) est donc essentiellement un outil d'annotation de textes permettant d'éditer un graphe de connaissances avec des informations sur ces deux niveaux d'abstraction. Les textes sont représentés de façon à permettre d'identifier des expressions d'arguments et de propositions et de décrire ces arguments et ces propositions. Une fois ces arguments et propositions identifiés, une autre partie de l'interface permet de créer des relations entre ceux-ci. Le résultat de l'utilisation de cet outil est donc un graphe de connaissances qui correspond à la lecture faite par l'utilisateur de ces textes. L'outil est développé de façon générique et peut donc être utilisé sur n'importe quel texte philosophique.

Un aspect particulièrement intéressant ici est de voir comment l'utilisation des technologies liées aux graphes de connaissances affecte la méthodologie de recherche utilisée et la façon de travailler sur l'étude de ces textes. En effet, la création du graphe de connaissances peut être vue comme une activité de recherche, et le graphe devient lui-même un objet d'étude, explorable et analysable, en plus des textes. De plus, cela facilite l'échange en rendant comparable les graphes de connaissances obtenus des lectures de différents utilisateurs. L'outil a d'ailleurs été testé dans un contexte d'enseignement, ou des étudiants en littérature et philosophie s'appuient sur l'annotation de textes comme moyen d'étude, et peuvent ensuite échanger sur leur compréhension de textes complexes sur la base de la comparaison des graphes obtenus.

6. Discussion : les avantages réels et les difficultés communes

Il n'est bien sûr pas dans l'intention de cet article de prétendre que les trois exemples de projets présentés ci-dessus sont représentatifs de projets typiques utilisant les graphes de connaissances pour les humanités numériques. Néanmoins, même s'ils ont de nombreux points communs, ceux-ci sont suffisamment différents pour que l'on puisse en tirer quelques leçons sur les avantages à utiliser des graphes de connaissances et certains problèmes qu'il reste à pallier.

En effet, un des éléments communs à ces trois projets est que les avantages obtenus ne sont pas nécessairement alignés avec ceux attendus. En effet, par exemple, seulement LED a été amené à développer une application intelligente et a réellement utilisé les liens avec d'autres sources d'information disponibles sous la forme de graphes de connaissances. Même pour LED, ces deux aspects ne sont pas réellement centraux et restent anecdotiques. De la même façon, l'accès à l'information reste important pour ces projets, mais sous une forme différente du

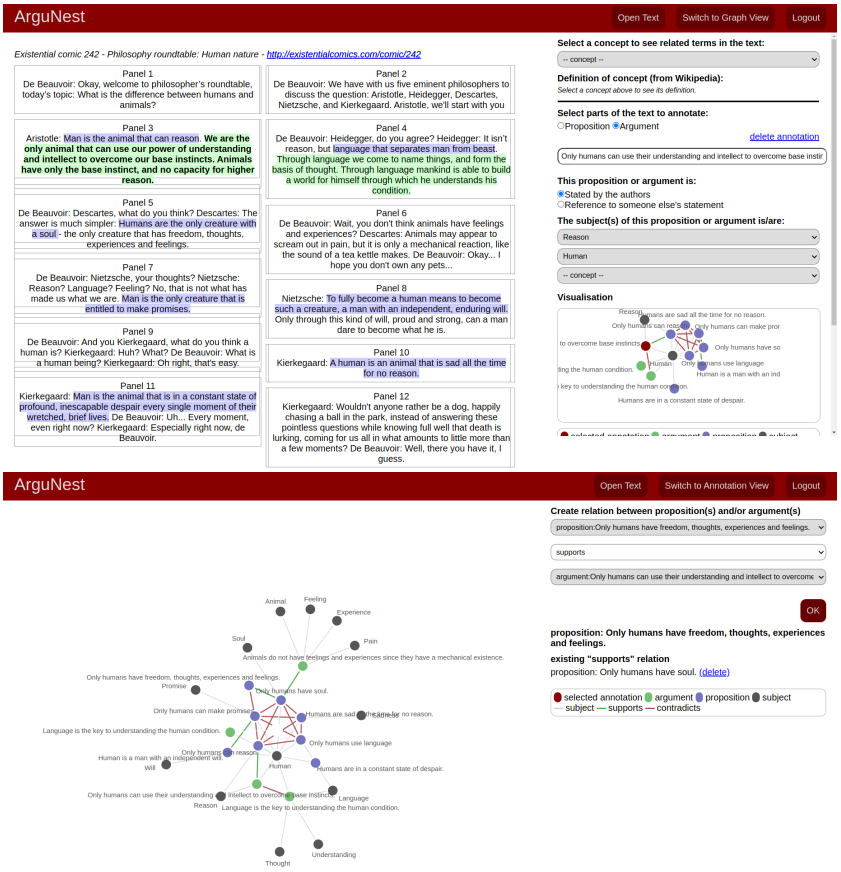

Figure 3 : L'outil ArguNest. Interface d'annotation de textes (haut) et de mise en relation des arguments et propositions (bas).

graphe de connaissances pur. Dans chacun de ces cas une interface masquant la technologie impliquée a été développée. Ces interfaces bénéficient de la mise sous forme de graphes de connaissances des informations à traiter, mais sont réalisées en plus pour faciliter l'exploration par des utilisateurs non familiarisés avec les principes du Web sémantique. Cela a, de fait, aussi une implication sur les attentes en termes de visibilité. D'autres projets autour du patrimoine culturel par exemple mettent cet aspect plus en avant, alors que dans le cas des trois projets considérés ici, l'apport de la création de graphes de connaissances à la visibilité du projet et de la recherche reste moindre.

Bien sûr, il ne faut pas conclure de ce qui est écrit ci-dessus que l'utilisation des technologies du Web sémantique n'a pas d'avantages. Tout projet en humanités numériques est, par nature, pluridisciplinaire, mais un des aspects essentiels de ces trois projets est qu'ils étaient tous les trois fondés sur une collaboration forte entre chercheurs en sciences humaines et chercheurs/développeurs en informatique. Comme montré plus haut, si on se focalise purement sur les éléments liés à la technologie, le fait que les graphes de connaissances permettent une

représentation flexible et évolutive semble être le point le plus important. En réalité, on peut aussi voir ce point technique comme étant fondamental au point positif le plus important dans ces trois projets : que la création sans contrainte technique forte d'une conceptualisation du domaine considéré représente une tâche pivot entre les disciplines, permettant d'en aligner les notions, les vocabulaires et les attentes.

En effet, dans chacun des trois projets, l'élément à la base de la collaboration et ayant permis de la construire était la création d'une ontologie qui réponde aux besoins du projet, et qui devait être le pilier central des outils et systèmes développés. La construction de cette ontologie doit, par nature, être une activité collaborative avec en son centre l'établissement d'une conceptualisation consensuelle des notions du domaine. Cette représentation doit être encodable dans les formalismes de représentation utilisés. Une des raisons de se tourner vers les technologies du Web sémantique et le graphe de connaissances est, comme déjà exprimé plus haut, qu'ils permettent des représentations riches, en incluant notamment des éléments incomplets et incertains. Les projets présentés ici ont débuté avec une vision vague et peu structurée de ce qui devait devenir cette conceptualisation. Là où d'autres types de technologies auraient nécessité de forcer une structure fortement contrainte sur la représentation des concepts du domaine et des informations associées dès le début du projet, la construction itérative d'une ontologie ici a permis une clarification progressive de ces éléments, l'explicitation des problèmes et des aspects spécifiques de chaque projet et la mise en place d'une vue partagée des éléments du projet au travers d'un dialogue plus équilibré entre le domaine de recherche concerné et la technique. Le résultat de ce processus est un artefact informatique qui non seulement va permettre de structurer le reste des développements technologiques dans le projet, mais qui va aussi encapsuler une vision commune du cœur du sujet entre les participants.

Que le plus significatif des avantages dans l'utilisation des graphes de connaissances soit au niveau de la conceptualisation et du dialogue entre les disciplines est aussi lié à un inconvénient majeur de ces technologies. En effet, malgré ce qui est écrit plus haut, les outils pour construire des ontologies, pour éditer des graphes de connaissances et pour naviguer au sein de ces graphes de connaissances restent complexes et obscures pour les non-spécialistes. Les conceptualisations initiales se font, par conséquent, souvent sur papier, sur des tableaux blancs ou sur la base d'outils non-dédiés, mais mieux maîtrisés ou maîtrisables par les experts du domaine de recherche. L'utilisation de tableurs en ligne dans le projet LEDA est un parfait exemple de ce type de difficultés qui nécessite de mettre en place des représentations intermédiaires que la partie plus technique du projet devra ensuite transformer en une représentation compatible avec l'utilisation des graphes de connaissances. Un autre exemple est le développement de l'outil OWBO [9] directement motivé par l'expérience dans ces trois projets. En effet, les outils de construction d'ontologies tels que Protégé [10] sont très peu adaptés à la phase initiale de structuration d'une ontologie et ne permettent pas facilement aux experts du domaine, non-spécialistes des technologies du Web sémantique, d'être directement impliqués dans la construction de l'ontologie. Une représentation séparée est souvent construite, par exemple sur un tableau blanc, laissant aux informaticiens le soin de les retranscrire dans Protégé. L'idée d'OWBO est

9. https://github.com/mdaquin/OWBO
10. https://protege.stanford.edu/

de fournir une version épurée, simplifiée et partageable de la création d'une ontologie initiale, qui s'apparente à l'utilisation d'un tableau blanc, et qui peut être transférée directement dans des outils tels que Protégé pour être affinée.

Finalement, un des désavantages les plus importants de l'utilisation des graphes de connaissances, peu visible dans la description des projets dans cet article mais tout de même très présent, est lié au manque de maturité des outils et des systèmes utilisés, et à leur pérennité. Beaucoup de ces systèmes sont développés dans des équipes de recherche ayant peu de moyens pour garantir leur fonctionnement et pour les mettre à jour autant qu'il peut être nécessaire. Comme discuté au début de cet article, alors que leur coût est moindre (ils sont souvent gratuits), le problème que cela amène est qu'il devient difficile de maintenir les graphes de connaissances et les applications construits sur la base de ces outils et systèmes. Le coût ici se retrouve concentré dans le temps des spécialistes en technologies du Web sémantique requis pour non seulement collaborer à la création de ces applications, mais aussi pour s'assurer que celles-ci continuent de fonctionner à long terme.

7. Conclusion

Dans cet article est décrit succinctement trois expériences de projets en humanités numériques utilisant des graphes de connaissances dans trois domaines différents : l'iconographie, l'histoire de la musique et la littérature/philosophie. Alors que pour les chercheurs impliqués dans le développement de ces technologies, les avantages de l'utilisation des graphes de connaissances peuvent paraître évidents, la réalité de leur mise en place et du développement d'applications les utilisant dans ce type de domaines n'est pas toujours alignée en pratique avec ce qui est attendu. Au travers de ces trois projets, il est possible de mieux comprendre comment, au-delà des aspects purement techniques, un des points essentiels des graphes de connaissances est qu'ils fournissent un modèle du domaine, complètement partagé et offrant une vision commune de ce qui est le cœur du projet. Néanmoins, pour que ces projets réussissent, il est nécessaire que cette conceptualisation soit réalisée au sein d'une vraie collaboration pluridisciplinaire entre les experts du domaine, qui possèdent la connaissance requise et les attentes liées au projet, et les spécialistes des technologies associées, capable de mettre en place les modèles de connaissances construits. Cela nécessite une forte implication de la part de ces spécialistes, au moment du développement et par la suite, d'autant que les outils disponibles actuellement ne sont pas vraiment conçus pour faciliter la collaboration ou la maintenance à long terme des graphes de connaissances et des applications les utilisant.

Remerciements

L'auteur tient à remercier les membres des projets LEDA, LED et ArguNest pour leur collaboration et contributions sur la base desquelles cet article a été écrit.

Références

[1] S. Ji, S. Pan, E. Cambria, P. Marttinen, S. Y. Philip, A survey on knowledge graphs : Representation, acquisition, and applications, IEEE Transactions on Neural Networks and Learning Systems (2021).

[2] G. Antoniou, F. Van Harmelen, A semantic web primer, MIT press, 2004.

[3] E. Hyvönen, Using the semantic web in digital humanities : Shift from data publishing to data-analysis and serendipitous knowledge discovery, Semantic Web 11 (2020) 187–193.

[4] S. Middle, Investigating Linked Data usability for Ancient World research, Ph.D. thesis, The Open University, Miton Keynes, UK, 2022.

[5] S. Staab, R. Studer, Handbook on ontologies, Springer Science & Business Media, 2010.

[6] D. Mouromtsev, M. d'Aquin, Open data for education : Linked, shared, and reusable data for teaching and learning, volume 9500, Springer, 2016.

[7] M. d'Aquin, E. Motta, The epistemology of intelligent semantic web systems, Synthesis Lectures on the Semantic Web : Theory and Technology 6 (2016) 1–88.

[8] R. Wenz, Linked open data for new library services : the example of data. bnf. fr., Linked open data for new library services : the example of data. bnf. fr. (2013) 403–416.

[9] A. Lymberopoulou, Hell in the Byzantine World : A History of Art and Religion in Venetian Crete and the Eastern Mediterranean, Cambridge University Press, 2020.

[10] A. Adamou, M. d'Aquin, H. Barlow, S. Brown, Led : curated and crowdsourced linked data on music listening experiences (2014).

[11] S. Brown, H. Barlow, A. Adamou, M. d'Aquin, The listening experience database project : Collating the responses of the" ordinary listener" to prompt new insights into musical experience., International Journal of the Humanities : Annual Review 13 (2015).

[12] A. Adamou, S. Brown, H. Barlow, C. Allocca, M. d'Aquin, Crowdsourcing linked data on listening experiences through reuse and enhancement of library data, International Journal on Digital Libraries 20 (2019) 61–79.