STRATEGE

Système de traçabilité et de gestion de l'information de données multimédia

R. Dib, J. Darmont, S. Loudcher, M. Pardoen

Le projet Bretez a pour objectif la restitution visuelle et sonore de la ville de Paris dans la seconde moitié du XVIIIe siècle. C'est un vaste projet pluridisciplinaire qui nécessite une collaboration entre différents domaines tels que les sciences humaines, l'histoire, la musicologie, la géographie, la sociologie et l'informatique. Il nécessite ainsi un important travail de recherche et de collecte d'informations provenant de musées et de bibliothèques sonores. Il est également indispensable de sauvegarder l'information et sa source, d'assurer son accessibilité afin de la vérifier et de documenter le travail. Ainsi, le besoin de traçabilité de l'information a donné naissance au module STRATEGE (Système de TRAçabiliTÉ et de GEstion de l'information de données multimédia).

Pour restituer le quartier parisien, les chercheurs du projet se sont appuyés sur des documents textuels, visuels et sonores des bibliothèques et des musées. Suite à l'identification des éléments constituants le quartier et à l'enregistrement des sons, un jeu vidéo a été développé avec le logiciel Unity qui permet de se déplacer virtuellement dans les rues, en entendant les différents bruits du quartier. Toutes les données collectées des différentes sources ont été enregistrées dans une base de données développée sous le logiciel FileMaker. Cette base de données servait au stockage et à la gestion des données. Toutefois, la variété des données (images, textes et sons) surpasse à l'heure actuelle la capacité de ce logiciel. Le défi augmente avec le développement du jeu vidéo et la nécessité de relier les éléments du jeu aux données qui leur correspondent dans la base de données, toujours dans le but de conserver la traçabilité de l'information. La question à se poser est alors : comment rassembler tous les textes, les images, les sons, les données de la base de données et du jeu vidéo, les rendre interopérables afin de garantir la traçabilité de l'information et de pouvoir les analyser plus tard ?

Devant l'hétérogénéité de tous ces éléments et le fait que certains soient structurés, telle que la base de données, et d'autres non structurés, tels que les images et les sons, nous avons opté pour un lac de données, un vaste espace de stockage de données hétérogènes provenant de sources de données externes, dans leur forme brute. Autrement dit, un lac de données permet la cohabitation de données structurées et non structurées du projet Bretez en respectant cette variété.

Ce rapport fait une synthèse des différentes étapes réalisées pour répondre à notre problématique. Dans la première partie nous présentons le concept d'un lac de données et nous justifions notre choix d'aborder ce concept en citant ses caractéristiques. Dans la deuxième partie nous résumons l'état des lieux des jeux de données et nous présentons les améliorations proposées. Finalement dans la troisième partie nous proposons une conception du lac de données pour le projet Bretez.

1. Lac de données

Introduit en 2010 par James Dixon, le concept de lac de données vise à réduire les limitations des entrepôts et des magasins de données. C'est un système évolutif de stockage et d'analyse de données de tous types, dans leur format natif, utilisé principalement par des spécialistes des données pour l'extraction de connaissances.

Tout d'abord, un lac de données stocke les données dans leur format brut. Grâce à l'approche *schema-on-read*, les données sont conservées dans leur format natif, sans opération de nettoyage. La structure des données n'est donc éventuellement copiée et modifiée que lorsqu'elles sont utilisées. Cela garantit ainsi le maintien de l'intégralité de l'information. Cette approche entre dans la démarche d'intégration des données *Extract-Load-Transform* (ELT). L'ELT ne garantit pas seulement la fidélité des données, mais permet également l'intégration des données en temps réel dans le lac. En effet, la différence entre le temps d'extraction des données et leur ingestion dans le lac est négligeable puisque les éventuelles transformations sont faites plus tard.

Par ailleurs, un lac de données permet une multitude des sources et l'hétérogénéité des données. C'est pourquoi il est caractérisé par une grande flexibilité et agilité. En effet, un lac de données est compatible avec tout type et format des données : qu'elles soient structurées, semi-structurées ou non structurées, les données cohabitent ensemble dans le lac qui respecte leur hétérogénéité.

De plus, un lac de données possède un système de métadonnées qui a pour rôle de décrire les données stockées pour permettre leur interrogation dans le lac. Il assure ainsi leur qualité. Les métadonnées sont divisées en trois parties : les métadonnées intra-objet qui indiquent les caractéristiques de chaque objet, telles que ses propriétés (titre, date de création…), sa définition métier et les modifications de versions et de représentations qu'il a subi ; les métadonnées inter-objets qui regroupent un certains nombre d'objets selon des caractéristiques précises ;et finalement les métadonnées globales qui servent à définir certains termes métier à l'aide d'un dictionnaire, par exemple. Un bon système de métadonnées est très important pour la gestion du lac et l'interrogation des données.

2. Jeux de données

Les jeux de données existants constituent la matière première pour la construction du lac de données. Nous disposons d'une base de données qui a servi de système de stockage des données pour le projet Bretez, de la maquette de jeu vidéo développé sous Unity et des documents images, textes et sons. Faire un état des lieux de ces données nous semble indispensable pour une bonne conception du lac de données.

En premier lieu, nous étudions la base de données. Cette base a été développée sous le logiciel FileMaker par les responsables, non informaticiens, du projet Bretez. La base de données est constitué de 7 tables principales et de 6 autres tables "ponts" qui font le lien entre les tables principales. Elle contient- les plans du quartier, ses bâtis et les références bibliographiques qui ont servi pour la documentation. De plus, cette structure stocke les images directement dans la base de données et cite leurs caractéristiques, ce que nous trouvons particulier et pas très efficace pour la gestion des données. C'est pourquoi nous avons proposé une deuxième version de la base de données implémentée sous MySQL. La nouvelle version enlève premièrement toutes les images de la base et les remplace par un lien vers ces fichiers. Deuxièmement, nous organisons la structure d'une manière d'avoir une même classe "Document" qui regroupe toutes les caractéristiques en commun entre les documents (images, textes, sons) et des sous-classes propre à chaque type. Ensuite nous nous sommes intéressés aux deux tables Plan et Censier. D'un point de vue métier ces deux tables représente des plans, mais les caractéristiques varient selon l'époque. Alors nous avons regroupé les deux dans une classe parent qui se divise en deux sous-classes pour garder en même temps le sens métier et la particularité de chacun. Finalement, nous effectuons des petites améliorations au niveau de la nomenclature des attributs l'enrichissement de certaines tables par des attributs essentiels.

En deuxième lieu, nous avons trois versions du jeu vidéo, la troisième étant la plus complète. Elles sont toutes développées sous le logiciel Unity. Comme l'objectif du projet Strategе est de relier les objets 3D du jeu vidéo aux éléments de la base de données, un travail d'extraction de ces objets est indispensable. Nous gardons ce travail pour une deuxième étape du projet qui nécessite des expertises de développement de jeu vidéo.

3. Conception du lac de données

Après la réalisation de l'état de l'art et de l'état des lieux nous nous intéressons dans cette partie à la conception du lac de données. Nous commençons par préciser les objets du lac et ensuite nous nous intéressons au système de métadonnées.

D'abord, nous devons conserver dans ce lac tous les fichiers images, textes et sons, la base de données initiale avec ses différentes représentations et sa nouvelle version et les trois versions du jeu vidéo.

Ensuite, d'après l'état des lieux effectués dans la partie précédente, nous remarquons que la base de données contient toutes les données métier ainsi que les données de propriétés des images, des textes et des sons. C'est ainsi que nous considérons cette base de données comme une base de métadonnées métier et multimédia et nous proposons de diviser le système de métadonnées en deux parties: les métadonnées métier et multimédia et les métadonnées du lac.

1. Métadonnées métier et multimédia:

Ce sont les métadonnées gérées par la base de métadonnées qui prennent la forme des attributs et des enregistrements dans cette base. Elles se divisent en deux parties intra-objet et inter-objets. Les métadonnées métier et multimédia intra-objet sont les propriétés des images, des textes et des sons ainsi que les métadonnées sémantiques de tous les objets du lac incluant les objets 3D du jeu vidéo. Les métadonnées métier et multimédia inter-objets sont les regroupements de tous les objets du lac selon les caractéristiques métier tel que l'appartenance au même plan ou au même bâti ou bien les regroupements des images, des textes ou des sons selon des caractéristiques spécifiques à chaque type.

2. Métadonnées du lac

Ce sont les métadonnées qui gèrent d'une part les métadonnées de propriétés des objets 3D du jeu vidéo et de l'autre part les métadonnées inter-objets des différents composants du lac. Les liens entre les images, les textes et les sont et la base de métadonnées sont considérés comme des métadonnées inter-objets logiques puisque la base de métadonnées conserve dans la table "Document" un attribut "lien" qui indique le chemin vers le fichier. Tandis que les liens entre la base de métadonnées et le jeu vidéo sont considérés comme des métadonnées inter-objets physiques. En effet, pour relier chaque objet 3D aux éléments qui lui correspondent dans la base de métadonnées nous implémentons une matrice de correspondance. Cette matrice associe l'identifiant de chaque objet 3D à l'identifiant de l'élément qui le décrit dans la base.

Finalement nous proposons un modèle conceptuel de notre lac de données qui suit le modèle de métadonnées MEDAL. Chaque objet du lac est représenté par un hyper-noeud qui contient des noeuds des versions et des représentations du même objet ainsi que des arcs qui indiquent ces transformations et mise à jour. Chaque noeud ou hyper-noeud possède ses propres métadonnées intra-objet. Les métadonnées inter-objets physiques sont représentées par des lignes joignant la base de métadonnées et le jeu vidéo à la matrice de correspondance. Les métadonnées intra-objets logiques sont représentées par des lignes pointillées joignant chaque fichier à la base de métadonnées. La figure 1 présente ce modèle conceptuel.

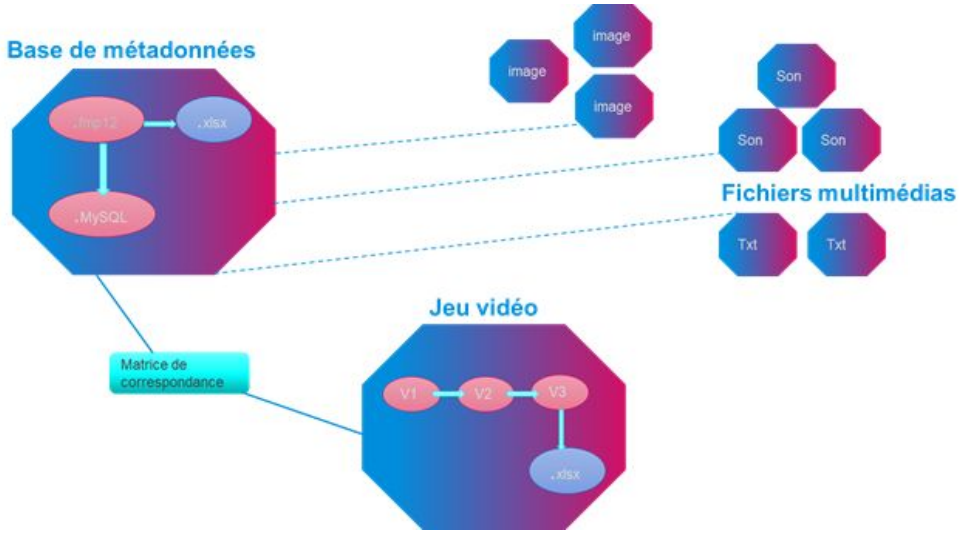

Figure 1: Modèle conceptuel du lac de données

Pour conclure, le projet STRATEGE vise à implémenter un lac de données pour permettre à des base de données, un jeu vidéo et des fichiers de types images, textes et sons à communiquer entre eux et d'être interrogeables afin d'assurer la traçabilité de l'information et l'analyse des données.

Le projet Strategе n'en est qu'à ses débuts. Nous avons renforcé la structure de la base de données et nous avons posé les fondations du lac de données. Les étapes suivantes seront de creuser plus profondément dans la maquette du jeu vidéo afin d'extraire ses objets, trouver les bonnes technologies pour implémenter le lac de données et finalement réaliser une interface graphique qui permet la bonne gestion du lac par les responsables, non *data scientists*, du projet Bretez.