La géomatique et la découverte de connaissance

L'analyse de données s'inscrit dans un processus de découverte de connaissance. La découverte de connaissance (Knowledge Discovery en anglais) est une extraction complexe à partir de la donnée d'une information implicite, au départ inconnue et potentiellement utile (Frawley *et al.*, 1992). Le processus de découverte de connaissance, explicité en Figure 1-1 est de nature itérative et permet de progresser depuis l'exploration des données jusqu'à la présentation de modèles. Ce processus en quatre phases aide l'usager à construire une connaissance dotée d'une composante spatiale (Gahegan *et al.*, 2001; Brisebois, 2003). La phase d'exploration (phase I) a pour objectif la recherche des patrons, de structures et de relations implicites dans un jeu de données. La phase suivante (II) consiste à proposer des hypothèses. Ces dernières sont par la suite validées ou invalidées au moyen d'analyses statistiques (phase III). Lorsque les résultats sont concluants (phase IV), ils sont communiqués en prenant soin de délimiter leur portée.

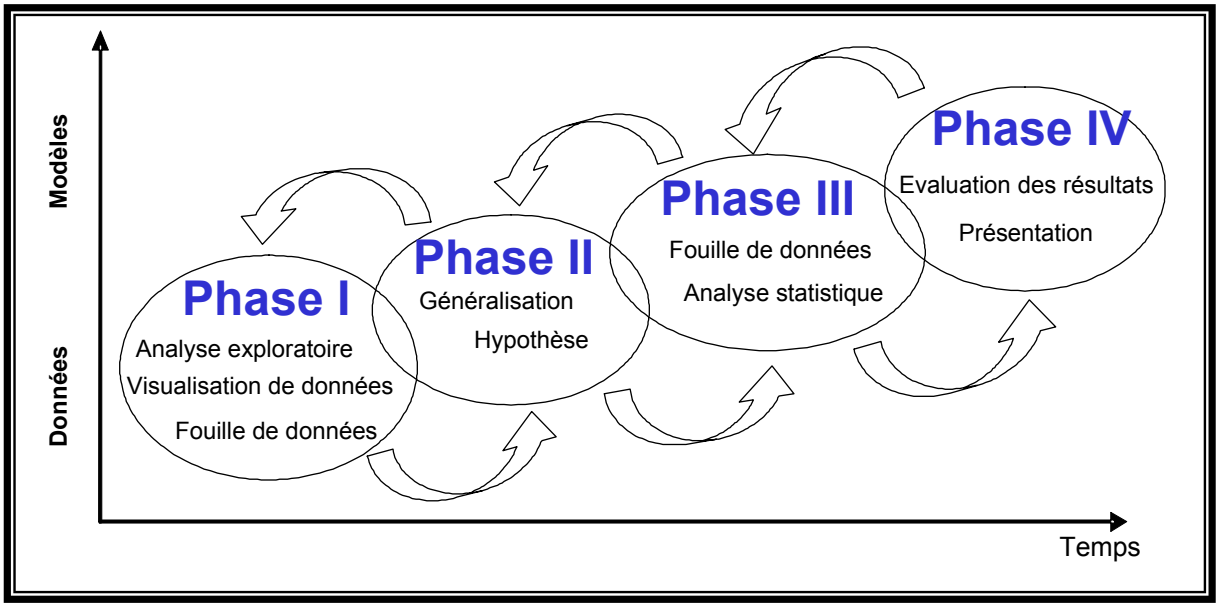


Figure 1-1 : Extrait du « Processus de construction de connaissances adapté de (Gahegan et al., 2001) » par (Brisebois, 2003)

La géomatique est une « discipline ayant pour objet la gestion des données géographiques et qui fait appel aux sciences et aux technologies reliées à leur acquisition, leur stockage, leur traitement et leur diffusion » (OQLF, 2006). Pour être compatible avec le processus de découverte de connaissance, la géomatique a dû inclure d'autres aspects : l'aide à la décision caractérisée, entre autres, par la technologie OLAP (On-Line Analytical Processing). Ces outils de découverte de connaissances ou d'analyse en ligne visent à assister l'usager dans son analyse en lui facilitant l'exploration de ses données et en lui donnant la possibilité de le faire rapidement (Bédard, 2005). Les travaux de ces dix dernières années sur l'analyse en ligne ont fait ressortir une perspective d'évolution intéressante. Ainsi, Caron, en 1998, a démontré que l'OLAP possède un potentiel réel pour supporter l'analyse spatio-temporelle (Caron, 1998). Cependant, sans volet cartographique, il est impossible de visualiser la composante géométrique des données spatiales. L'ajout de ce volet cartographique a conduit à l'élaboration de l'outil SOLAP (OLAP spatial) : « une plate-forme visuelle supportant l'exploration et l'analyse spatiotemporelle rapides des données selon une approche multidimensionnelle à plusieurs niveaux d'agrégation » (Rivest, 2000). Par la suite, la potentialité de la tridimensionnalité de la technologie SOLAP a été explorée (Brisebois, 2003). Suite aux constats de Brisebois et aux difficultés à manipuler la tridimensionnalité des données (Zlatanova *et al.*, 2002; Lachance, 2005), l'archéologie est apparue comme un excellent domaine d'application. Elle offre des situations permettant de se pencher sur la manipulation des données tridimensionnelles dans un contexte de découverte de connaissance.

1.1-2. L'opportunité archéologique : l'analyse spatio-temporelle de données 3D

Lors de la fouille d'un site archéologique, l'archéologue/fouilleur prend en compte la localisation sur le terrain (x, y, z) des données qui y sont découvertes. Suivant une pratique courante, elles sont enregistrées en regard d' « unités de fouille » (**UF**) qui se trouvent être des volumes (tridimensionnels) de terre archéologique qui sont retirés de différents endroits du site lors des opérations de fouille et qui contiennent des artéfacts, des écofacts, des restes humains ou des vestiges architecturaux.

Lors du processus d'analyse, les « unités de fouille », et par conséquent le matériel archéologique qu'elles contiennent, sont mises en relation avec les différentes couches de terre stratigraphiques formant le site. Ces corrélations tiennent une place importante dans le processus de découverte de connaissances archéologiques. En effet, la compréhension - ou l'interprétation - d'un site archéologique passe, dans un premier temps, par la reconstitution de l'ordre d'accumulation au sol de ces couches stratigraphiques. Puis, dans un second temps, grâce à des jalons chronologiques fournis par le matériel archéologique, la séquence de déposition des couches est datée d'une manière relative – les unes par rapport aux autres - mais aussi absolue : points fixes dans le temps. A noter que les vestiges archéologiques enregistrés sur un chantier de fouilles sont des traces laissées au sol par des humains dans le passé. Le but ultime du fouilleur est de reconstituer la séquence événementielle qui a mené à la formation du site archéologique qu'il a fouillé : « Un site archéologique est un lieu où, dans le passé, des actions humaines et sociales ont été réalisées » (Barceló *et al.*, 2003).

Les méthodes les plus répandues à l'heure actuelle parmi les archéologues de terrain pour procéder à la construction (ou reconstruction) des connaissances spatio-temporelles sur un chantier de fouilles sont la combinaison des dessins en plan et en coupe, ainsi que la « Matrice de Harris ». Cependant, ces moyens « traditionnels » agissent dans un environnement 2D et non 3D comme le souhaiteraient les archéologues dans leur optique de reconstruire une réalité 3D (Green *et al.*, 2001; Malinverni *et al.*, 2002; Barceló et al., 2003; Cattani *et al.*, 2004; Day *et al.*, 2004; Losier, 2005).

Des catégories de logiciels gèrent déjà des données géographiques 3D : ce sont les outils de types Conception Assistée par Ordinateur (plus communément appelés CAO ou CAD en anglais) et les Systèmes d'Information Géographique (plus communément appelés SIG ou GIS en anglais). Plus spécifiquement, de nombreux travaux ont été réalisés dans le domaine des SIG 3D en urbanisme (De La Losa, 2000; Ramos, 2003) ou en géologie (Apel, 2004). Cependant, les technologies actuelles en logiciel SIG ne permettent pas de répondre à la problématique particulière de l'archéologie citée précédemment (Wheatley *et al.*, 2002; Zlatanova *et al.*, 2004; Barceló, 2005). En effet, jusqu'à présent, l'évolution de la recherche s'est plus orientée vers la modélisation et la navigation 3D au détriment d'une analyse de

données 3D (Zlatanova et al., 2004). Ainsi, force est de constater que, pour l'instant, tous les outils permettant la gestion des données tridimensionnelles ont plus mis l'emphase sur la visualisation et la représentation 3D que sur une analyse de données (descriptive, temporelle, spatiale ET visuelle) rapide et simple.

Cette situation lance des défis à la géomatique notamment la nécessité d'analyser simultanément l'espace et le temps, dans un contexte où l'un peut être influencé par l'autre, et de le faire rapidement et facilement dans un environnement tridimensionnel afin de mieux supporter l'interprétation des données, la création d'hypothèses et la découverte de nouvelle connaissance archéologique.

1.2-Problématique

Initialement, notre intérêt s'était porté vers une amélioration des fonctionnalités tridimensionnelles des outils SOLAP dans un domaine – l'archéologie de terrain – où l'on connaît l'existence d'une problématique d'analyse tridimensionnelle spatio-temporelle. Même si la pertinence d'un outil SOLAP 3D en la matière a déjà été démontrée (Brisebois, 2003), l'efficacité réelle d'un tel outil pour les utilisateurs n'a pas été testée en profondeur puisqu'il s'agissait d'une première exploration technologique du potentiel d'adapter le SOLAP au 3D ainsi que d'une première tentative d'amener un système de type SOLAP en archéologie (Rageul, 2004). L'amélioration des fonctionnalités tridimensionnelles devait passer par une étude de la gestion des données spatio-temporelles tridimensionnelles dans un contexte SOLAP. La mise en évidence des variables graphiques utiles au système d'analyse de données 3D, l'intégration d'un modèle géométrique volumique aux systèmes SOLAP étaient des exemples d'amélioration. Nous devions nous servir de l'opportunité archéologique afin d'étudier la faisabilité et l'efficacité de tels outils. Plus particulièrement, l'étude de leur efficacité devait permettre aux utilisateurs potentiels d'évaluer l'outil en fonction de sa facilité à extraire l'information des données descriptives, temporelles et spatiales issues du processus d'acquisition des données et de permettre une utilisation et une rapidité d'exécution visuelle compatible avec le processus analytique des archéologues.

Cependant, le cours de recherche a permis de mieux comprendre les attentes des archéologues quant à l'analyse de leurs données. « Un problème dans la réalisation de projets SOLAP est la difficulté des utilisateurs à exprimer leurs besoins et leurs attentes en début d'analyse de système ». (Guimond, 2005). Ainsi l'analyse des attentes des archéologues a permis la découverte d'un autre besoin, plus immédiat et intéressant tout autant pour eux que pour la géomatique. La découverte de ce besoin a permis de réorienter la problématique initiale vers une nouvelle problématique tout en offrant certains éléments de réponse à la première. Ce nouveau besoin n'a jamais été rencontré auparavant et il présente un défi primordial pour l'archéologue : **la possibilité de modifier les données incluses dans un SOLAP « durant » l'analyse**. Cette modification n'entre ni dans la mise à jour des données terrain, ni dans l'optique d'une correction d'erreur mais plutôt dans une nouvelle catégorie de problème. Dans la Figure 1-1 relative aux phases de découverte de nouvelles connaissances, nous pouvons situer ce besoin principalement dans la boucle de rétroaction entre les phases 1 (analyse exploratoire) et 2 (génération d'hypothèse).

Il existe en effet des situations exceptionnelles (comme en archéologie) où la donnée incluse dans l'outil SOLAP doit être « modifiée », « **révisée** », i.e. réinterprétée afin d'être améliorée. Il s'agit d'une amélioration que l'utilisateur veut apporter à la donnée car en fait, et c'est le cas en archéologie, l'analyste ne peut valider efficacement le peuplement de l'outil SOLAP qu'en effectuant son analyse avec l'outil en question ! En d'autres termes, c'est en explorant ses données qu'il fait émerger des idées, de nouvelles hypothèses, et c'est grâce à celles-ci qu'il peut améliorer son interprétation du site et qu'il « construit » la réalité d'autrefois. Dans l'outil SOLAP, la qualité des données s'améliore donc au fur et à mesure de son analyse et ce sont justement ces données qui fournissent la clé aux archéologues.

Or, actuellement, les systèmes OLAP servent principalement à exploiter une base de données ne subissant qu'occasionnellement des ajouts (ce qui est différent des modifications) et dont la structure est aussi rarement modifiée. En raison d'une dénormalisation volontaire, pour satisfaire une réponse rapide à des requêtes plus ou moins complexes, les outils OLAP traditionnels sont plutôt inadaptés pour exploiter des données

subissant des ajouts fréquents (Bédard, 2005). Les systèmes OLAP « temps réels » ont justement été développés pour répondre spécifiquement à ce besoin. Cependant ces systèmes ne sont pas encore optimisés à la mise à jour complète (ajout, suppression et modification) des données ce qui se ferait normalement suite à un changement sur le terrain dans la réalité. En effet, le processus de mise à jour consiste à changer la valeur des données pour les faire correspondre le plus possible à l'état actuel de la réalité (d'où le terme également utilisé d'actualisation des données) (Pouliot *et al.*, 2004). Les systèmes OLAP « temps réels » ne peuvent également pas forcément permettre la révision interactive de la structure du système et des données avec des modifications des caractéristiques de celles-ci. La possibilité de pouvoir modifier le contenu du SOLAP pendant une phase d'analyse interactive, voire même la structure complète de l'outil, ne semble pas avoir été explorée à ce jour dans le monde des données spatiales. De plus, les concepts de mise à jour de données ont été enrichis, mais dans le sens de rafraîchissement (ajout) de données suite à la reconstruction de système pour un entrepôt de données (Lambert, 2005). A la différence du projet de Lambert, notre problématique de recherche consiste à **étudier l'évolution, qui plus est interactive, des données qui en plus sont tridimensionnelles, dans une structure SOLAP**. Cette évolution ne concerne pas exclusivement le rafraîchissement et la reconstruction des données, mais aussi une révision possible des caractéristiques d'une donnée ou de la structure du système SOLAP, tout en laissant à l'utilisateur, l'opportunité de le faire pendant sa démarche de découverte de connaissance.

1.3-Objectifs

L'objectif principal de notre recherche est de **proposer des mécanismes qui permettraient l'évolution des données 3D spatio-temporelles (ajout, suppression et modification) dans une structure multidimensionnelle** directement durant la phase d'analyse de l'utilisateur.

De cet objectif principal découlent des objectifs secondaires :

- Faire état des réflexions actuelles sur les possibilités de combiner l'exploration 3D avec les solutions OLAP et/ou SOLAP afin de faire des recommandations pour le développement futur de ce genre de système.

- Elargir les concepts d'évolution de données et de structure dans une structure multidimensionnelle
- Identifier les besoins des archéologues quant à l'analyse de leurs données, mettre en évidence leur processus analytique et faire des propositions pour mettre au point un système de découverte de connaissance mieux adapté à leur contexte.

1.4-Méthodologie

La méthodologie suivie au cours de la présente recherche est exposée en Figure 1-2.

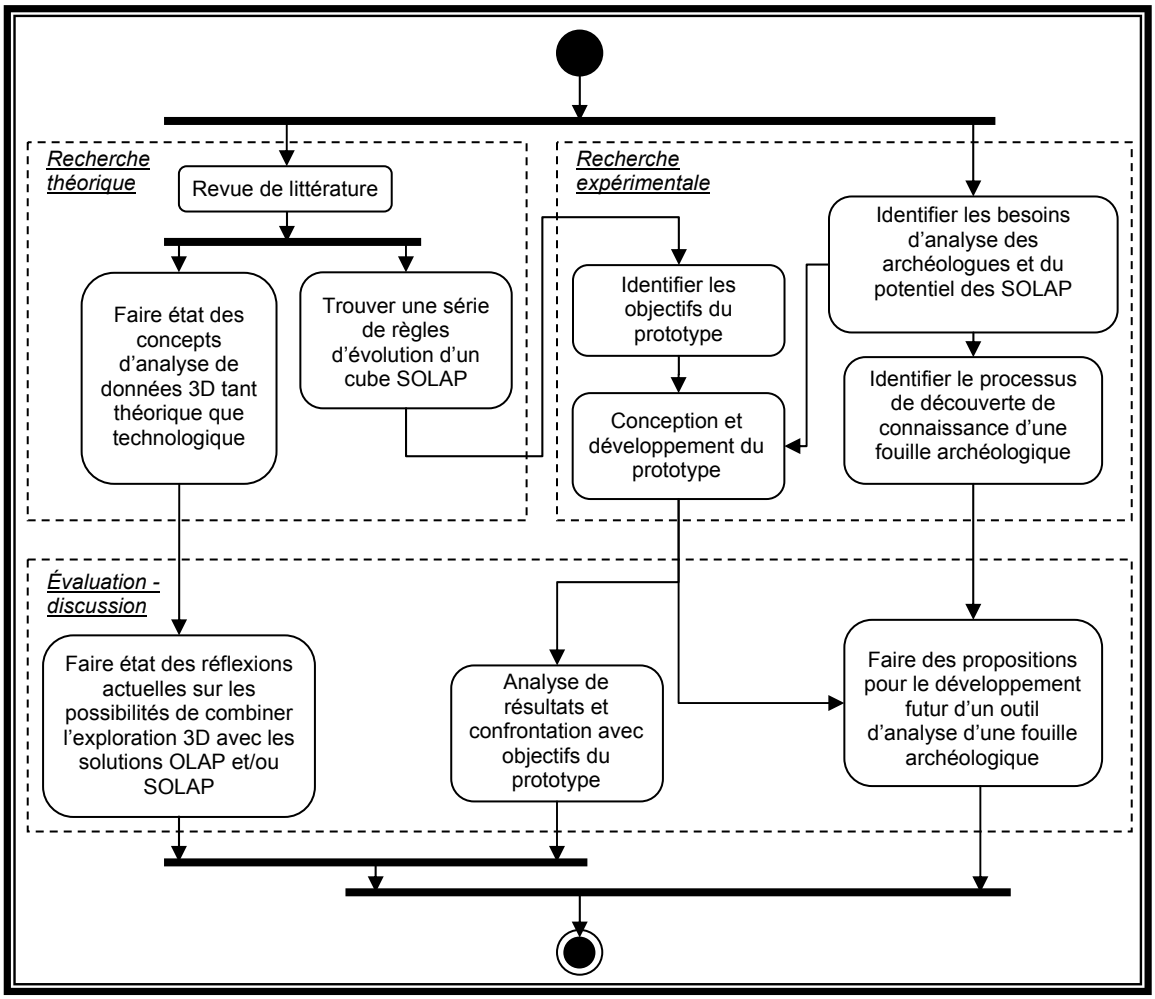


Figure 1-2 : Méthodologie adoptée dans la présente recherche (formalisme UML : diagrammes d'activités)

La *recherche théorique* s'est limitée à faire état des concepts d'analyse de données 3D tant théorique que technologique. La revue de littérature s'est portée sur l'analyse de données

graphiques et descriptives, sur les différents types de technologies à potentiel 3D : SIG 3D, CAO 3D entre autres, mais aussi sur les diffusions de données graphiques telles que la réalité virtuelle comme outil de visualisation 3D. Cette recherche concernait aussi la mise en place d'une série de règles d'évolution d'un cube SOLAP.

La *recherche expérimentale* s'est portée tout d'abord sur l'identification des besoins des archéologues. Diverses rencontres avec des archéologues ont permis d'identifier une partie du processus de découverte de connaissance d'une fouille archéologique. Ces rencontres ont également permis d'étudier le potentiel du SOLAP appliqué à l'archéologie à l'aide d'une maquette SOLAP résultante du projet de géomatisation du processus de fouille archéologique (Rageul, 2004) et présentée à des archéologues. La recherche expérimentale a aussi conduit à la mise en place d'un prototype d'évolution de données d'un cube SOLAP. Le projet étant appliqué à l'archéologie, l'étude préalable des besoins d'analyse des archéologues a été nécessaire.

L'*évaluation et la discussion concluent* la recherche sur les trois thèmes abordés au cours de cette maîtrise :

- la mise en place d'architectures justifiées combinant l'exploration 3D avec les solutions OLAP et/ou SOLAP,
- l'évolution des données et de la structure pendant la phase d'analyse dans une structure multidimensionnelle
- le contexte archéologique (identification des besoins d'analyse, exploration du processus de découverte de connaissance d'une fouille archéologique et recommandations pour le développement futur d'un outil d'analyse adapté à leurs besoins.

1.5- Description des chapitres du mémoire

Ce mémoire est divisé en 5 chapitres. Après la présente introduction (chapitre 1), le chapitre 2 expose une synthèse des notions théoriques pertinentes à notre projet de recherche. Cette synthèse caractérise la donnée spatiale tridimensionnelle, en examinant quelles sont les catégories d'analyses effectuées et les technologies disponibles pour gérer, manipuler, analyser et visualiser cette donnée. Nous y avons introduit les concepts du SOLAP et de l'approche multidimensionnelle et plus particulièrement les dernières optimisations (temps réel et 3D) relatives à notre projet. Le chapitre 3 décrit une approche théorique d'optimisations des outils SOLAP. La première optimisation concerne la prise en compte de l'évolution de données pendant la phase d'analyse. La seconde optimisation porte sur les architectures potentielles d'un SOLAP 3D et comprend une discussion sur les difficultés à les mettre en œuvre. Le chapitre 4 explicite, quant à lui, l'approche expérimentale de la première optimisation dans le contexte archéologique. Nous nous attacherons aussi dans ce chapitre à comprendre le processus de découverte de connaissance archéologique. Enfin, le chapitre 5 conclut ce mémoire en discutant les résultats de nos approches tant théorique qu'expérimentale, et en offrant des pistes de recherche qui pourraient être explorées dans le cadre d'autres projets de recherche qui seraient susceptibles de voir le jour au regard de nos conclusions.

Chapitre 2 : Revue des concepts

2.1- Introduction

Afin d'introduire certains concepts théoriques jugés nécessaires à la compréhension de notre travail de recherche, nous allons brièvement discuter sur les données et leur rôle prépondérant dans la représentation numérique 3D. Ce chapitre introduit également les concepts d'analyse, processus permettant de déduire l'information des données pour mieux comprendre la réalité. Ces notions sont importantes car elles permettent de poser les bases tant dans notre approche théorique d'optimisation (chapitre 3) que notre approche expérimentale (chapitre 4).

Il nous a également semblé pertinent de faire une présentation des différents systèmes informatiques capables d'exploiter les données, et tout particulièrement celles qui sont tridimensionnelles. Ces systèmes informatiques tiennent une place importante dans la démarche de l'analyste lorsque celui-ci est dans une phase de compréhension de la réalité. Nous mettrons l'accent sur le système SOLAP ainsi que deux optimisations : la première concerne la notion de temps réel dans un SOLAP et la deuxième concerne la gestion de la donnée tridimensionnelle. Nous verrons par la suite dans le chapitre suivant comment ces deux optimisations influence directement notre recherche d'optimisation.

2.2- La donnée : de sa création à la génération de modèles 3D

Comme le souligne Aamodt et Nygard, la donnée, l'information et la connaissance sont trois notions extrêmement dépendantes (Aamodt *et al.*, 1995). Les Figures 1-1 et 2-1 représentent le processus de construction de connaissances, la donnée servant de base à la génération de modèle et à l'élaboration de nouvelle connaissance. La **donnée**, en géomatique, est la représentation d'une information codée dans un format permettant son traitement par ordinateur (OQLF, 2006). La donnée peut être un nombre, un texte ou un symbole [(Longley et al., 2005) p 11] non interprétés (Davenport, 1997; Van der Spek *et al.*, 1997). Elle est physique et peut donc être stockée. Lorsque nous donnons un sens, une signification, une interprétation spécifique à une donnée, nous la transformons en **Information** (Davenport, 1997; Van der Spek et Spijkervet, 1997; Choo *et al.*, 2000;

Pouliot, 2005). Dès qu'on décode les données dans leur contexte d'utilisation et selon la personne qui les lit, une même donnée peut recevoir diverses interprétations et peut donc conduire à différentes informations (ou bien à aucune information si on n'en comprend ni le caractère, ni le symbole). L'information est un élément de connaissance concernant un phénomène et qui, pris dans un contexte déterminé, a une signification particulière. Le cadre de référence qui détermine cette interprétation est constitué de la somme des **connaissances** (ensemble de faits, d'événements, de règles d'inférence et d'heuristiques) et des expériences de la personne qui effectue l'interprétation (OQLF, 2006).

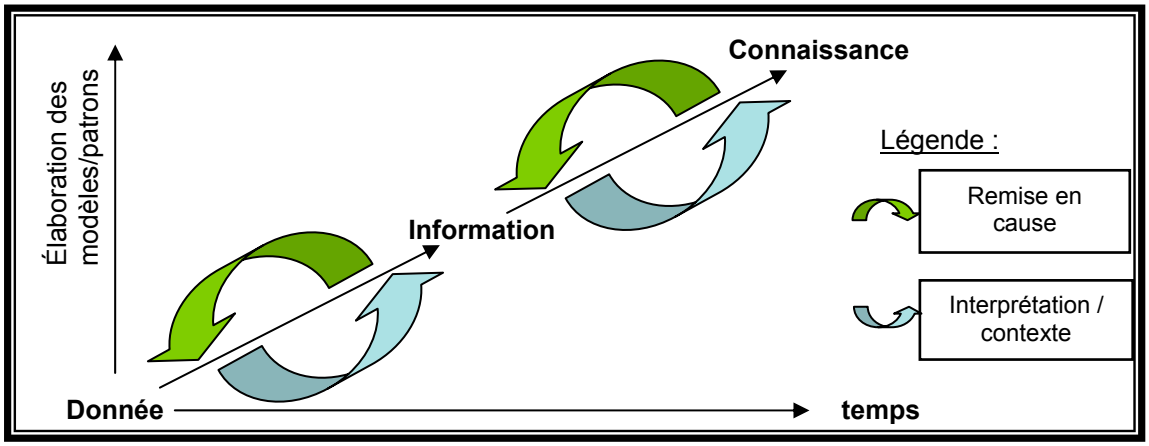


Figure 2-1 : De la donnée à la connaissance inspiré de (Bellinger *et al.*, 1997; Gahegan et al., 2001)

Afin de comprendre comment la donnée pourra nous informer dans les divers systèmes de découverte de connaissance, nous allons montrer la terminologie utilisée pour passer de la réalité à sa représentation numérique tridimensionnelle.

2.2-1. De la réalité aux différents types de données

Le « Robert » définit la réalité comme étant constituée de « choses réelles et de faits réels », c'est à dire qui « existent en fait ». En géomatique, la réalité est constituée de phénomènes et de relations entre ces phénomènes. L'OQLF (2006) définit le phénomène comme la réalité qui se manifeste à la conscience, que ce soit par l'intermédiaire des sens ou non. Le phénomène constitue donc la réalité première. Comme le montre graphiquement la Figure 2-2, le phénomène peut prendre l'apparence d'un objet tangible comme une maison, d'un événement comme un accident, d'un concept comme la notion d'unité de fouille en archéologie ou finalement l'apparence d'un être vivant comme une personne. L'autre

élément constitutif de la réalité est la relation entre les phénomènes, comme par exemple : « Monsieur Jalon possède une maison ».

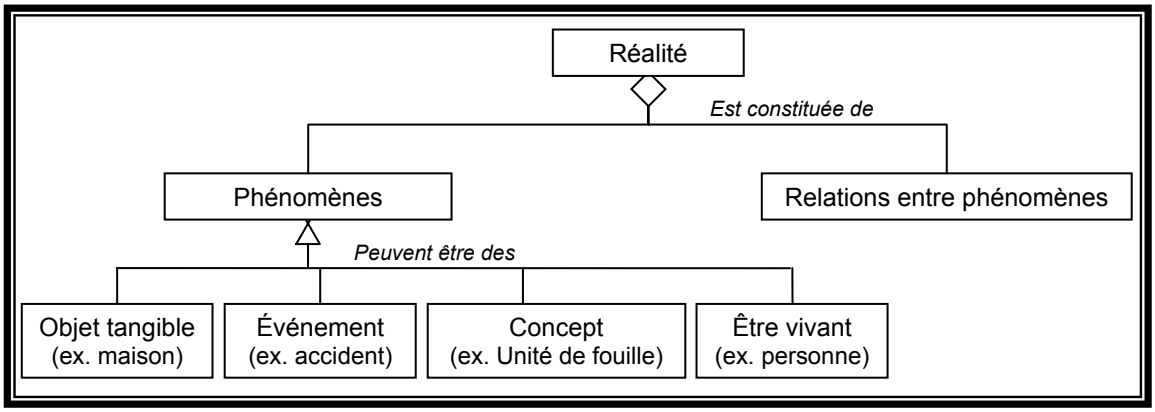


Figure 2-2 : La caractérisation d'une réalité inspirée de (Bédard, 2003a)

Afin de comprendre et de représenter numériquement cette réalité, nous allons chercher à la modéliser. La modélisation est une description dans un langage compréhensible à la fois par l'humain et par l'ordinateur de la forme, du mouvement et des caractéristiques d'un objet ou d'un ensemble d'objets qui crée un modèle, c'est-à-dire une représentation simplifiée de l'objet (OQLF, 2006). Cependant, un modèle reste une abstraction spécifique de la réalité et il doit faire ressortir de manière explicite les caractéristiques de ce qu'il est sensé modéliser. C'est en effet sur le modèle que l'analyse se fait d'où un contrôle permanent de la validité du modèle comme étant une représentation conforme à la réalité (Wilsey, 1999).

Étant donnée que les systèmes d'analyses visés par notre recherche sont principalement constitués de base de données, il n'est pas inutile de définir la démarche de modélisation ; démarche nécessaire pour passer d'une réalité à sa représentation : « Une méthode de modélisation est un ensemble de procédures et de règles à suivre permettant de capter la

réalité du client et de concevoir une base de données correspondant aux besoins de cette réalité » (Bédard, 2003a). Au niveau conceptuel, c'est à dire indépendamment de la technologie utilisée, les modèles[1] sont constitués de classes d'objets - ou d'entités suivant la méthode utilisée - décrites par des attributs, des opérations, des associations entre les classes d'objets, des agrégations de classes d'objets simples en classes d'objets complexes et des généralisations ou spécialisations d'une super-classe en ses sous-classes. Les phénomènes sont ainsi modélisés en classes d'objets. Cet objet est l'élément de base qui sert à construire des logiciels. Il est la matérialisation de la classe dont il reproduit les caractéristiques (OQLF, 2006).[2] Les relations entre phénomènes, quant à elles, sont modélisées par des associations qui peuvent être de surcroît spatiales et/ou temporelles et décrites dans les classes d'objets (cf. Figure 2-3).

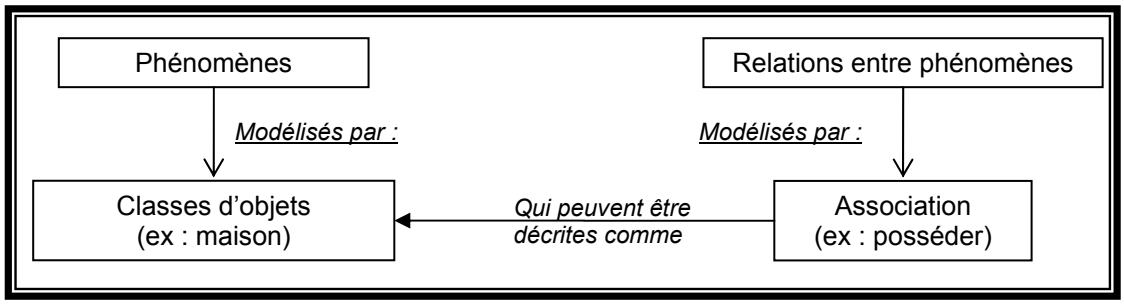


Figure 2-3 : La modélisation de la réalité en classe d'objets inspiré de (Bédard, 2003a)

Ainsi, les classes d'objets (ou entités), éléments essentiels d'un modèle conceptuel de données (MCD), sont décrites par des attributs et des opérations (cf. Figure 2-4). L'attribut est une « composante d'un modèle conceptuel de données, qui représente une caractéristique propre à un phénomène ou une caractéristique propre à une relation entre phénomènes » (OQLF, 2006). On y retrouve les attributs descriptifs comme le nombre d'étages d'un bâtiment, les attributs temporels comme la date de construction du même bâtiment, les attributs géométriques comme la position en (x,y,z) et les attributs graphiques comme la couleur. La donnée se retrouve ainsi stockée dans les valeurs d'attribut.

[1] Modèle conceptuel de données pour la méthode « Entité/Relationnel » ou modèle d'analyse pour la méthode « Orienté Objet »

[2] Cette notion d'objet se différencie de l' « objet tangible » présent en Figure 2-2. A partir de la section 2.2-2. la notion d'objet sera considéré comme un objet tangible, ie que l'on peut connaître en touchant.

Les **données descriptives** (cf. Figure 2-4) sont des données relatives à un des attributs d'une entité ou d'une relation, à l'exclusion de sa position et de sa forme (OQLF, 2006). Cela comprend les **attributs descriptifs** comme le nombre d'étages d'un bâtiment par exemple, mais aussi les **attributs temporels** comme la naissance ou la mort de l'objet. On peut aussi rencontrer dans la littérature la notion de **données temporelles** comme étant une donnée qui peut avoir un lien avec toute notion de temps (durée ou instant) (Korte, 2001).

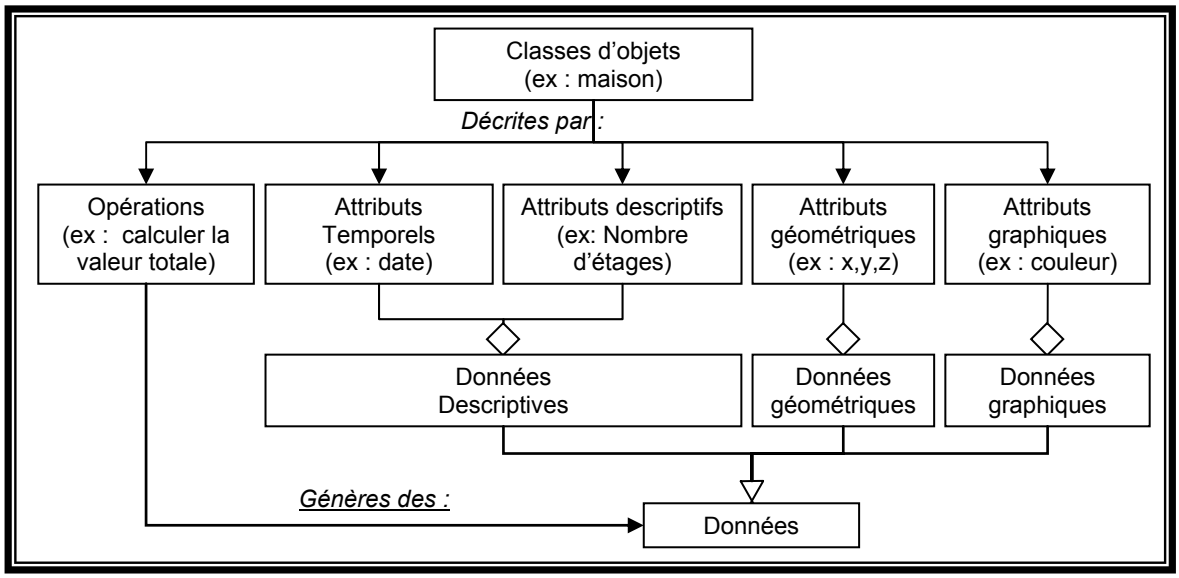


Figure 2-4 : des classes d'objets aux données inspiré de (Bédard, 2003a)

Les **données géométriques** (cf. Figure 2-4) renseignent sur la position ou la forme d'une entité géométrique. Les primitives géométriques (point, ligne, polygone) et les primitives topologiques (nœud, vecteur, surface) en sont des exemples (OQLF, 2006).

Les **données graphiques** (cf. Figure 2-4) portent sur la représentation visuelle d'une entité géométrique On peut ainsi mentionner, comme exemple, la couleur et la largeur du trait. On parle de données graphiques pour les représentations des entités spatiales sur des photos, des cartes ou des plans : la représentation cartographique des limites d'une propriété par exemple (OQLF, 2006).

Les **données spatiales** portent sur les entités spatiales et leurs relations, dans une application géomatique. L'objet peut alors être localisé par rapport aux autres [(Haining, 2003) : p4]. Les données spatiales comprennent l'ensemble des données géométriques, des

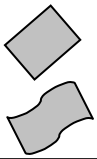
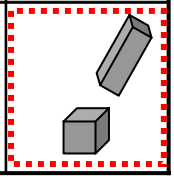
données descriptives et des métadonnées utilisées dans l'application géomatique (OQLF, 2006). Les données spatiales archéologiques étant des données spatiales tridimensionnelles, nous avons tenté de trouver une définition d'une **donnée spatiale 3D**. Cependant, celle-ci n'existe pas de manière **explicite** et **uniforme** dans la littérature principalement à cause de divergences liées à la définition même de « 3 » et « D » (Larrivée *et al.*, 2005). La section qui suit a pour but de présenter le rôle de la donnée pour la représentation numérique 3D.

2.2-2. L'objet tridimensionnel

L'objet est un « phénomène quelconque du monde extérieur ou intérieur qu'un homme observe (ou peut observer) à un instant déterminé » (OQLF, 2006). Il peut être décrit selon diverses dimensions : géométriques (en lien avec les données géométriques) ou descriptives (en lien avec les données descriptives). La notion de dimension est utilisée dans différents domaines (dessin technique, économie, électricité, imprimerie, mathématique, philosophie, informatique, physique, statistique…). Cependant, elle n'est pas clairement définie (Brisebois, 2003; Lachance, 2005; Larrivée *et al.*, 2006; Pouliot *et al.*, 2006) Face à cette multitude de définitions, nous utiliserons deux notions de dimensions liées à l'objet multidimensionnel (espace) pour définir l'objet 3D tel que nous l'utiliserons pour ce mémoire. Ce sont la dimension géométrique d'objet en lui-même et la dimension de l'univers dans lequel l'objet est positionné.

La **dimension géométrique de l'objet** est exprimée en fonction du nombre de directions ou de lignes suivant lesquelles le corps s'étend (Brisebois, 2003). D'après (Bédard *et al.*, 2002), l'objet 3D (encadré pointillé dans le Tableau 2-1) sera associé à une dimension géométrique 3D, c'est à dire un *objet volumique.*

Tableau 2-1 : La dimension géométrique des objets adaptée de Brisebois 2003

	Géométrie 0D	Géométrie 1D	Géométrie 2D	Géométrie 3D
Exemples d'objets				

La **dimension de l'univers** dans lequel évolue l'objet est exprimée en fonction du nombre d'axes, spatiale et/ou temporelle, nécessaire et suffisant pour positionner l'objet. D'après (Kennedy, 2004), l'objet 3D (encadré pointillé dans le Figure 2-5) sera associé à une dimension de l'univers 3D, c'est à dire à trois axes spatiaux et où les objets sont positionnés dans le volume infini et imaginaire, formé par les 3 axes.

Dépendant de la littérature étudiée, un objet peut donc être considéré comme 3D (Brisebois, 2003; Lachance, 2005; Pouliot, 2005) si la géométrie de l'objet est une géométrie 3D (Bédard *et al.*, 2002) ou si l'univers dans lequel évolue l'objet est un univers 3D (Kennedy, 2004) (Figure 2-5) ou encore si l'enveloppe des objets, c'est-à-dire le rectangle englobant de l'objet, est en 3D (Bentley, 2002) (cf. Figure 2-6).

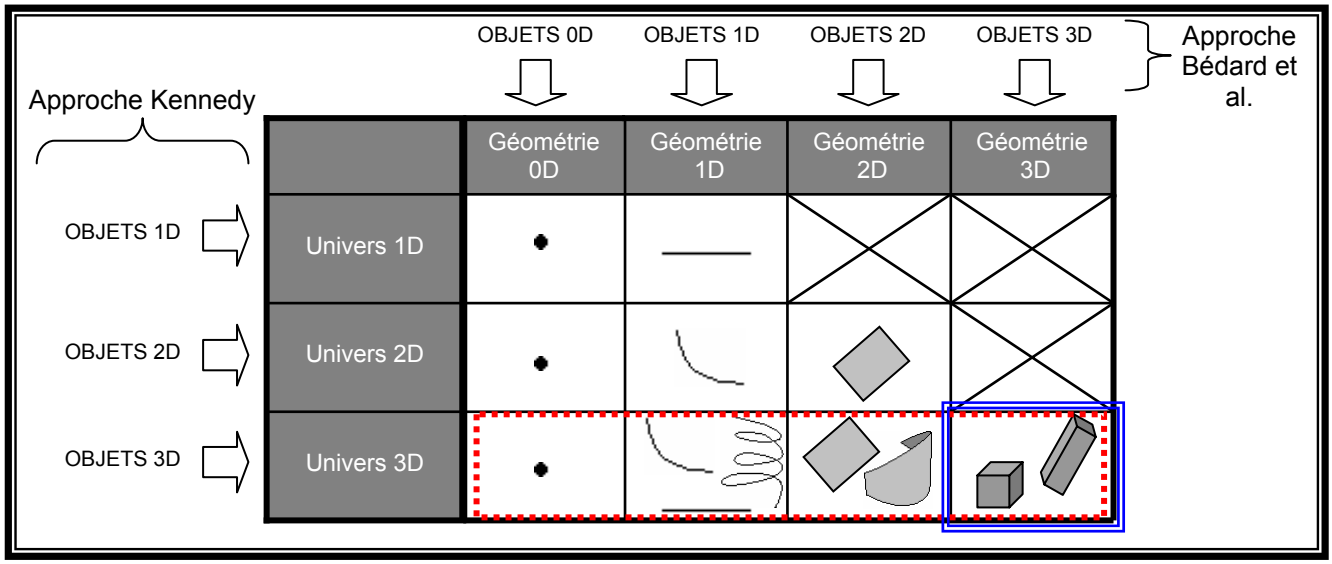


Figure 2-5 : Comparaison de l'objet 3D défini par la géométrie et défini par l'univers adapté de (Brisebois, 2003)

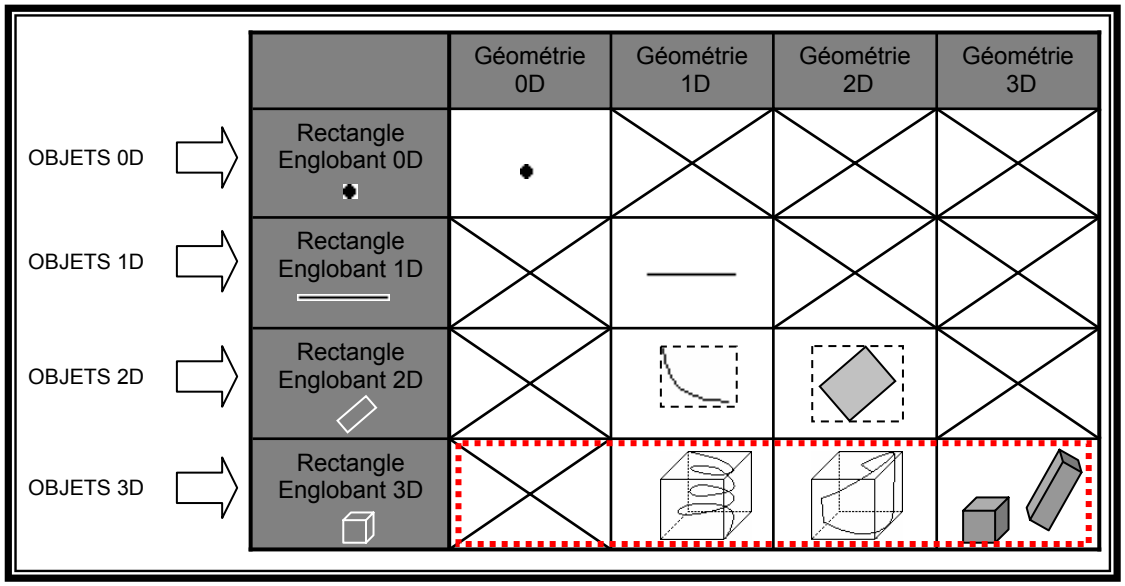


Figure 2-6 : Objet 3D défini par son rectangle englobant adapté de (Brisebois, 2003)

Pour le présent mémoire, nous associerons le terme d'objet 3D à **tout objet à trois dimensions géométriques (longueur, largeur et hauteur) et positionné dans un univers 3D (x, y, z) et référencés temporellement** (encadré double sur la Figure 2-5). En effet, en archéologie, lors des opérations de fouille, les seules données spatiales sont enregistrées en regard d' « unités de fouille » (*UF*) qui se trouvent être des volumes (tridimensionnels) de terre archéologique qui sont retirés de différents endroits du site (univers 3D).

2.2-3. La modélisation géométrique 3D

Les Figure 2-2, Figure 2-3 et Figure 2-4 ont montré le cheminement qu'il faudrait suivre pour reconstituer la réalité. À partir des données, la forme, le mouvement et les caractéristiques des objets sont reconstitués sous forme de modèle. Étant donné que nos travaux chercheront à examiner tout particulièrement l'évolution des informations dans un contexte d'analyse spatiale 3D et parce que l'archéologie propose déjà des modèles géométriques diversifiés (Green et al., 2001; Barceló et al., 2003; Nigro *et al.*, 2003; Cattani et al., 2004; Day et al., 2004; Mngumi *et al.*, 2004; Losier *et al.*, 2007) il nous apparaît d'expliquer en quoi consiste la modélisation géométrique tridimensionnelle.

La modélisation tridimensionnelle rend compte « des propriétés géométriques de l'objet » (OQLF, 2006) et s'intéresse à la représentation de la géométrie des systèmes étudiés et s'interroge sur la représentation spatiale proprement dite et sur la manière de les représenter (Pouliot, 2005). Cependant, il n'existe pas une seule façon de construire un modèle géométrique [(Longley et al., 2005) : p178]. Cette différence marque l'importance de bien spécifier quel type de modélisation géométrique est utilisé, puisque la modélisation influence de près l'analyse des données et l'affichage de celles-ci (Pouliot, 2005). Beaucoup d'auteurs (De La Losa, 2000; Ramos, 2003; Apel, 2004; Lachance, 2005) ayant déjà fait un état de l'art sur les différentes modélisations, l'intérêt de notre synthèse se porte plutôt sur une classification, même si elle ne se prétend pas exhaustive, qui permettrait de

mieux cerner les avantages et les inconvénients de chacune. À ce sujet, nous avons retenu la même classification que celle utilisée par Pouliot qui propose une classification basée sur les approches orientées espace et orientées objets (Pouliot et al., 2006).

L'**approche orientée espace** est utilisée lorsque le partitionnement de l'espace est arbitraire (par exemple des pixels ou des triangles) (Pouliot et al., 2006). Cette approche est souvent associée à la structure matricielle ou raster. D'après Longley, le modèle de données raster utilise des cellules (par exemple des pixels en 2D ou voxel en 3D) pour représenter les objets du monde réel [(Longley et al., 2005) : p181]. Il existe plusieurs manières de découper l'espace. La première, le découpage régulier, se fait sous la forme de cubes élémentaires fixes ou variables : le voxel. Ce découpage permet une représentation non ambiguë, unique et simple de la réalité (Pouliot, 2005; Thalmann, 2006). La précision, la granularité du modèle dépend directement de la résolution du pixel/voxel ou du nombre de niveaux hiérarchiques de l'octree. Cependant, cette précision se fait au détriment d'une taille de plus en plus conséquente du fichier de stockage. La deuxième manière, le découpage irrégulier, permet d'avoir une meilleure résolution avec un meilleur stockage. Le meilleur rapport résolution/stockage est principalement dû au fait que le découpage peut être lié à l'orientation de certains objets (modèles Binary Space Partitioning) ou en fonction de la densité des points échantillonnés (modèles en décomposition en cellules).

L'**approche orientée objet** est utilisée lorsque le partitionnement de l'espace est fonctionnel (par exemple la frontière d'une route, d'un bâtiment) (Pouliot et al., 2006). Une prémisse importante de cette approche consiste à connaître les frontières de l'objet. Il faut donc avoir ces frontières et être capable de les mesurer et de les estimer. Cette approche est souvent associée à la structure vectorielle. Les objets sont construits à partir des coordonnées des points et des arrêtes décrivant leurs position et forme. Il existe plusieurs approches pour représenter de façon discrète l'enveloppe des objets 3D ou leurs parties. L'*approche basée sur la frontière* est caractérisée par les modèles « fils de fer » (ou wireframe en anglais) et les modèles B-rep (Boundary Representation) La modélisation *fil de fer* est la plus simple des modélisations 3D [(Bertoline *et al.*, 2002) : p 304] car elle ne contient que les informations sur les points (vertex) et les lignes /courbes (edge) [(Foley,

1995) : p560]. Cette simplicité dans la modélisation permettant de stocker les informations géométriques la rend performante mais en fait aussi un modèle ambigu, confus et non-unique [(Saxena *et al.*, 2005) : p258 et 259]. La modélisation *B-rep* (boundary representation), modélise les faces des objets (orientées ou non) pour représenter les solides. [(Bertoline et Wiebe, 2002) : p314] sans pour autant avoir une idée sur ses propriétés volumiques (De La Losa, 2000). L*'approche basée sur des formes paramétrables* est caractérisée par les modèles Constructive Solid Geometry (plus communément appelés modèles CSG), les modèles par primitive Instancing et les modèles par balayage. La construction des objets avec la méthode *CSG* consiste à l'élaboration d'un modèle géométrique complexe à partir d'un jeu de primitives simples (Skibniewski *et al.*, 1997) comme des cubes, des sphères, des cylindres, des cônes ou des tores [(Gasparini, 2005) : p63] qui sont emboités après translation, rotation et facteurs d'échelle puis auxquels on effectue des unions, des intersections et/ou des différences (De La Losa, 2000). La primitive instancing est une approche indépendante de la représentation des objets solides. Cette modélisation est basée sur la notion de famille d'objets ou chaque membre de la famille est distingué par des paramètres [(Foley, 1995) : p539](Skibniewski et Kunigahalli, 1997) et ne se chevauche pas entre eux (Lattuada, 2005). Finalement, la modélisation par balayage consiste à représenter un objet en « balayant », une aire définie ou un volume, le long d'une trajectoire définie (Lattuada, 2005).

2.3-L'analyse de données – revues de définitions

Afin de comprendre la réalité et d'améliorer sa perception du monde, les analystes essaient de reconstruire cette réalité sous forme de modèle. L'élaboration de modèles se fait à l'aide des différentes données recueillies. Ces données sont caractérisées en fonction de leurs positions spatiale et temporelle, de leur description et de leur représentation graphique. L'aspect spatial est d'autant plus important puisqu'il sert de base à la modélisation géométrique. L'archéologie est une discipline exploitant au maximum cette reconstruction de la réalité au moyen de modèles géométriques (Losier, 2005). Une fois les données et le processus de modélisation décrits, nous allons maintenant expliquer les principaux types d'analyse à disposition de l'analyste (cf. chapitre3). Cela nous permettra de mieux mettre

en évidence les tenants et aboutissants d'un tel processus dans le contexte spécifique de l'analyse des données issues d'un chantier de fouille archéologique (cf. chapitre 4).

L'analyse de données correspond à l'ensemble des méthodes statistiques permettant de visualiser, de classer et d'expliquer les données (OQLF, 2006). Il s'agit d'un processus cognitif itératif qui utilise différents opérateurs et qui vise à mieux faire connaître un phénomène et à le situer dans un plus grand ensemble, afin d'améliorer notre modèle cognitif de la réalité (cf. Figure 2-7) (Champoux, 1991). Ce processus cognitif permet à l'analyste de déduire l'information qu'il désire obtenir à partir des données présentées par son outil d'analyse (Caron, 1998).

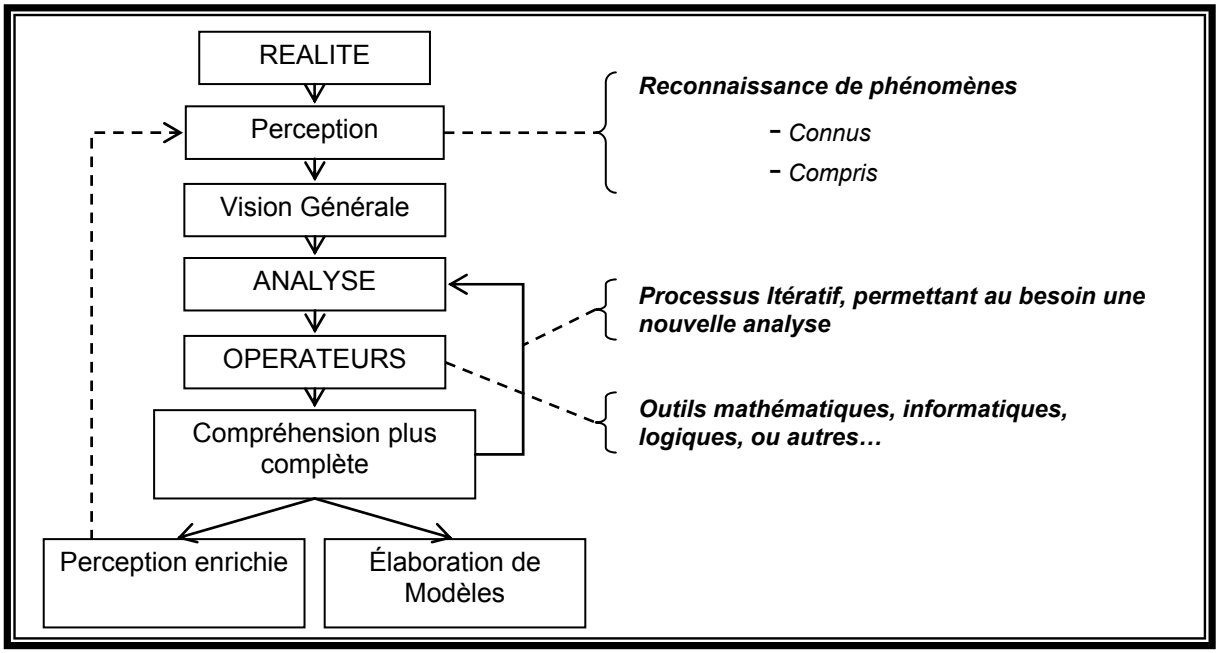


Figure 2-7 : L'analyse, de la réalité à la compréhension, adaptée de (Champoux, 1991)

2.3-1. Les différentes catégories d'analyses

L'**analyse spatiale** est un terme couramment utilisé dans les SIG. En 1987, Goodchild a défini l'analyse spatiale comme un ensemble de méthodes analytiques qui requiert l'accès à la localisation ainsi qu'aux attributs des objets analysés (Goodchild, 1987). En 1989, Aronoff complète cette définition en ajoutant que les fonctions d'analyse spatiale utilisent les attributs spatiaux et non spatiaux pour répondre à des questions sur le monde réel (Aronoff, 1989). En 1991, Champoux, définit l'analyse spatiale comme un « processus cognitif et itératif qui utilise différents opérateurs, dont au moins un spatial pour déduire

des caractéristiques descriptives ou spatiales d'un phénomène isolé ou regroupé, réel ou simulé dans l'espace. » (Champoux, 1991). Finalement, en 1995, Proulx précise l'analyse spatiale comme étant un « processus cognitif et itératif qui utilise différents opérateurs, dont *au moins un spatial, mais aucun temporel*, pour déduire les caractéristiques descriptives, spatiales ou temporelles d'un phénomène isolé ou regroupé, réel ou simulé dans l'espace » (Proulx, 1995). Cette dernière définition étant celle qui convient le mieux à l'archéologie, considérant leur nécessité à différencier l'analyse spatiale de l'analyse temporelle, nous retiendrons donc la définition de Proulx pour ce mémoire.

Parallèlement à l'analyse spatiale d'un objet qui porte sur la partie géométrique, l'**analyse temporelle** porte sur la datation, sur l'existence ou non, sur l'évolution ou non, de l'objet. L'analyse temporelle est une analyse de données en relation avec le temps. Cette analyse est très souvent utilisée par exemple dans le domaine de la criminalité [(Mena, 2002) : p47], mais aussi sur des modèles de simulation, sur des suivis de trafics routiers. Nous utiliserons la définition de Proulx pour définir l'analyse temporelle. L'analyse temporelle est un processus cognitif et itératif qui utilise différents opérateurs, *dont au moins un temporel, mais aucun spatial*, pour déduire les caractéristiques descriptives, spatiales ou temporelles d'un phénomène isolé ou regroupé, réel ou simulé dans le temps (Proulx, 1995).

Nous utiliserons aussi la définition de Proulx pour expliciter l'**analyse spatio-temporelle**. L'analyse spatio-temporelle est un processus cognitif et itératif qui utilise différents opérateurs, *dont au moins un temporel et au moins un spatial*, pour déduire les caractéristiques descriptives, spatiales ou temporelles d'un phénomène isolé ou regroupé, réel ou simulé dans l'espace et dans le temps (Proulx, 1995). Cette analyse spatio-temporelle est particulièrement importante en archéologie. Elle permet d'analyser simultanément l'espace et le temps, dans un contexte où l'un peut être influencé par l'autre.

L'**analyse descriptive** est d'abord un moyen de faire ressortir des unités statistiques. En effet, cette analyse a pour objectif de résumer quantitativement ou qualitativement l'information récoltée sur l'échantillon de données étudié. L'explication n'est alors pas

recherchée ; l'objectif est seulement de décrire, en dégageant les faits essentiels (Juillard, 2005; Sanaa, 2005). Traditionnellement, la donnée descriptive tient compte non seulement du contenu descriptif d'une donnée mais aussi du contenu temporel (Proulx, 1995; Korte, 2001; ESRI, 2006). L'archéologie nécessitant de traiter tant l'aspect spatial que temporel de la donnée, il est impérieux de différencier la temporalité de l'aspect descriptif des données. La notion de données descriptives existant dans la littérature, elle nous amène à une définition de l'analyse descriptive évidente : *analyse qui porterait sur les données descriptives* et donc qui ne porterait ni sur l'aspect spatial, ni sur l'aspect temporel de la donnée. Cependant, malgré une révision exhaustive de la littérature, nous n'avons trouvé aucune définition explicite sur l'analyse descriptive pour les bases de données autre que celle-ci. « Une analyse descriptive est un processus cognitif et itératif qui utilise les opérateurs descriptifs pour déduire les caractéristiques descriptives d'un phénomène » (Proulx, 1995). Par exemple, l'entité spatiale municipalité possède certains attributs descriptifs: nom, population, activité économique, etc. Il est donc possible d'effectuer une analyse descriptive du genre "Quelles sont les villes de plus de 100 000 habitants?" (Boursier, P., Mainguenaud, M., 1992).

La dernière catégorie d'analyse que nous allons présenter est **l'analyse visuelle ou graphique**. L'homme est meilleur pour interpréter visuellement des données que pour interpréter des chiffres [(Longley et al., 2005) : p273]. De plus, l'image tient une place importante dans la réflexion, l'interprétation et la compréhension car quelque soit la représentation (diagrammes, icones, géométries,…), ce sont des symboles qui ressemblent plus à ce qu'ils sont censés représentés que le texte. (Thagard *et al.*, 1997; Kovalerchuk *et al.*, 2005). L'expérience de générations de scientifiques comme Bohr, Einstein, Faraday et Watt ont montré que la représentation et le raisonnement visuels peuvent grandement améliorer la recherche de nouvelles hypothèses [(Kovalerchuk et Schwing, 2005) : p74]. L'analyse visuelle s'appuie sur ce que Bertin appelle la Graphique [(Bertin, 1992) : p20]. Elle poursuit deux objectifs : traiter les données pour comprendre et en tirer l'information et communiquer s'il y a lieu cette information ou inventaire de données élémentaires. Ces deux objectifs sont ce que Bertin appelle la Graphique de traitement et la Graphique de communication. La *Graphique de traitement* comporte deux impératifs auxquels la

graphique de communication n'est pas soumise: elle doit transcrire toutes les données du tableau, c'est-à-dire les données exhaustives et elle doit répondre à toutes les questions pertinentes et permettre de simplifier les deux composantes du tableau de données. La *Graphique de communication* est un moyen de fixer et de dire aux autres ce que l'on a découvert d'une manière simple et rapide. Elle utilise les propriétés de l'image visuelle pour faire apparaître les relations de ressemblance et d'ordre entre les données. Elle permet alors une perception rapide et éventuellement la mémorisation de l'information d'ensemble.

2.3-2. Les opérateurs d'analyse:

Les opérateurs sont des outils mathématiques, informatiques, logiques ou autres qui permettent à l'utilisateur d'effectuer des requêtes (Champoux, 1991). Un requête est un ensemble de commandes dont l'exécution permet d'obtenir un résultat (OQLF, 2006). Plus spécifiquement, un opérateur est l'élément qui exécute un calcul (Ginguay *et al.*, 1993) lors d'analyses descriptive, spatiale, temporelle ou spatio-temporelle (Proulx, 1995). A cette liste d'opérateurs, on peut rajouter les opérateurs d'analyse graphique ou visuelle ce qui permet alors de distinguer 4 grandes familles d'opérateurs : les opérateurs spatiaux, les opérateurs temporels, les opérateurs descriptifs et les opérateurs visuels.

Les opérateurs spatiaux

D'après l'OQLF (2006), un opérateur spatial est un opérateur permettant de traiter la position et la forme d'une entité spatiale, ou les relations spatiales qui existent entre les entités spatiales. Que l'on travaille avec des données 2D ou des données 3D, il existe deux grandes familles d'opérateurs spatiaux : (Champoux, 1991; Bédard, 2003b) : les opérateurs spatiaux métriques et les opérateurs spatiaux topologiques.

Les **opérateurs spatiaux métriques 2D ou 3D** sont des opérateurs qui permettent de traiter tout ce qui se mesure (Golay *et al.*, 2001). D'une part, on y retrouve les simples mesures comme une distance, une direction, une moyenne, un calcul de périmètre, de surface, de volume,… D'autre part, on y retrouve aussi tout ce qui peut toucher à l'analyse de position complexe notamment les rayons d'inclusion, les corridors, les enveloppes,…

Les **opérateurs spatiaux topologiques 2D et 3D** permettent de traiter les relations spatiales établies entre des entités spatiales (OQLF, 2006). Ceux-ci sont utilisés pour établir la présence ou l'absence de certaines relations spatiales entre les entités géométriques. (Champoux, 1991) La topologie est une branche des mathématiques traitant des relations de voisinage entre des figures géométriques, relations qui ne sont pas altérées par la déformation des figures (OQLF, 2006). Cela signifie que même après des transformations comme une translation, une rotation ou un changement d'échelle, la propriété topologique des figures restent invariante (Lachance, 2005). Kemp complète en spécifiant que dans le contexte d'une structure vectorielle, la topologie est la description explicite de relations spatiales entre des entités stockées (Kemp, 2006). Dans les SIG, la topologie peut être utilisée pour valider la géométrie vectorielle (c'est à dire, vérifier qu'un des polygones se ferme et que toutes les lignes du réseau se ferment ensemble) et pour certains types d'opérations (analyse réseau et test d'adjacence) [(Longley et al., 2005) : p186].

Il existe différentes manières de définir le nombre de relations topologiques entre des objets (Pullar *et al.*, 1988; Egenhoffer *et al.*, 1991; Clementini *et al.*, 1993). D'après Normand et al., le modèle reconnu par la norme ISO TC 211 – norme ISO 19107 sur le schéma spatial (www.iso.org) et basée sur le concept d'Egenhofer (Egenhoffer et Herring, 1991; Egenhoffer *et al.*, 1992; Egenhoffer *et al.*, 1994) – peut être regroupé en 5 catégories (Normand et al., 2003). La **disjonction** est utilisée lorsqu'il n'y aucun contact entre l'intérieur et la limite des objets. L'**adjacence** est utilisée lorsqu'il y a contact avec que les limites des objets. L'**intersection** est utilisée lorsqu'il y a contact entre l'intérieur d'un objet et la limite avec (intersection intérieure) ou sans (intersection limite) l'intérieur des autres objets. L'**inclusion** est utilisée lorsqu'un objet contient complètement un autre objet, avec une limite commune (inclusion limite) ou non (inclusion totale). L'**égalité** est utilisée lorsque les deux objets se recouvrent exactement.

Le logiciel *ArcMap* d'ESRI (www.esri.com) avec les opérateurs topologiques « Intersect, are completely within, completely contain, touch the boundary of, …» ou MGE d'intergraph (www.intergraph.com) avec les opérateurs « overlaps, entirely contains, passes trough, meets, Contained by, on boundary of,… » ou le logiciel Map Info

(www.mapinfo.com) avec les opérateurs « contains, within, contains entire, contains part, intersect,… » sont autant de preuves de la présence des logiciels actuels à gérer les opérateurs spatiaux topologique 2D. En 3D, les opérateurs topologiques, en théorie, sont relativement similaires avec le modèle en 9 intersections d'Egenhofer. Cependant, ils sont bien plus difficiles à mettre en place puisqu'ils exigent une structure topologique 3D de l'objet, plus complexe qu'une structure topologique 2D (De La Losa, 2000; Zlatanova, 2000; Zlatanova et al., 2002; Billen *et al.*, 2003; Ramos, 2003; Apel, 2004; Zlatanova et al., 2004; Lachance, 2005; Pouliot, 2005; Pouliot et al., 2006).

Les opérateurs temporels

Pour que des relations temporelles puissent être considérées entre deux objets, il faut supposer le temps T comme quelque chose de continu. L'objet temporel est alors considéré dans ce temps définit. Il est caractérisé par un intervalle temporel I définit par deux instants t1 et t2 (t1<t2) (Claramunt *et al.*, 2000). Les opérateurs temporels sont alors basés sur le même principe que les opérateurs spatiaux. [(Swiaczny *et al.*, 2001) : p139]. On peut donc étendre la classe opérateurs métriques et topologiques aux requêtes temporelles (Gagnon, 1993; Bédard, 2003b).

Gagnon (1993) les **opérateurs métriques temporels** comme la position d'un objet dans le temps ou la durée de celui-ci. A l'instar des opérateurs spatiaux topologiques, les opérateurs temporels topologiques concernent également la relation immuable, au niveau temporel, entre les objets. Les concepts d'Allen (Allen, 1984) sont souvent utilisés pour définir les relations topologiques temporels (equal, before after, meets, overlaps, during, starts, finishes) et leurs inverses respectifs à l'exception d' « equal » qui est un opérateur symétrique (after, met, overlapped, contain, started, finished) (cf. Figure 2-8) (Claramunt et Jiang, 2000; Jian-feng *et al.*, 2005).

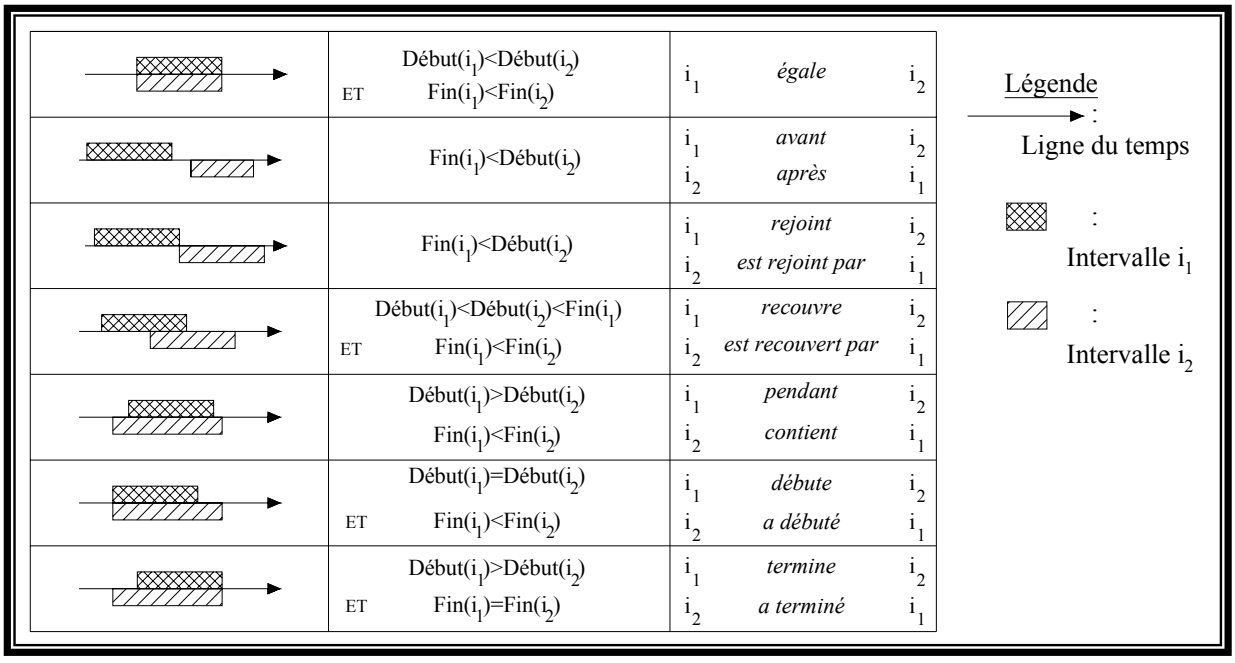


Figure 2-8 : Les opérateurs temporels topologiques

Les opérateurs descriptifs

Proulx (1995) a défini les opérateurs descriptifs comme des opérateurs utilisant des attributs descriptifs de la base de données comme critères de sélection.

On distingue deux catégories d'opérateurs descriptifs (Proulx, 1995; Bédard, 2003b) :

- les opérateurs mathématiques qui sont appliqués exclusivement aux valeurs quantitatives. On dit d'une valeur qu'elle est quantitative si cette valeur peut-être mesurée ou comptée (OQLF, 2006). Ces valeurs, numériques seront par exemple de type *float*, *double* ou *integer* (Diansheng, 2003). Les opérateurs « +, -, *, / » sont des exemples d'opérateurs mathématiques
- les opérateurs logiques qui sont appliqués aux valeurs qualitatives mais aussi quantitatives. Les valeurs qualitatives peuvent être soit de nature nominale (les valeurs sont des noms dont l'ordre n'a pas d'incidence comme *commercial, résidentiel, industriel*) ou de nature ordinale (les valeurs sont des noms qui suivent une progression comme *pauvre, moyen riche*) (Bédard, 2003a; Pouliot, 2003). Ces valeurs, alphabétiques, seront de types *char* ou *varchar* (Diansheng, 2003). Les opérateurs « <, >, =, <=, >=, <> » sont des exemples d'opérateurs logiques. Les valeurs qualitatives de nature nominale utiliseront exclusivement les opérateurs logiques « = » et « <> ». De plus, d'après Proulx (1995), les opérateurs logiques comprennent les opérateurs ensemblistes (ex: "and", "or").

Les fonctions de visualisation de l'information spatiale

L'interprétation visuelle des données peut se faire à l'aide d'opérateurs ou des fonctions de visualisation de l'information spatiale (Golay et Caloz, 2001). La liste d'opérateurs visuels ne peut être ni exhaustive ni hiérarchisée (du plus important au moins important) en raison de l'aspect subjectif du sens de la vue ; chaque personne ayant sa propre perception et son propre sens de l'analyse visuelle. Cependant, on peut retenir les principales fonctions de visualisation suivantes :

- l'élaboration de cartes thématiques – avec l'importance de la légende et de l'orientation
- la visualisation tridimensionnelle – il est estimé que 50 % des neurones humains sont réservé à la vision et que l'affichage 3D stimule plus de neurones, utilisant

ainsi une plus grande portion du cerveau permettant la résolution plus rapide et facile des problèmes (Swanson, 1996).

- la navigation – zoom et déplacement de l'écran (cf. Figure 2-9)
- l'utilisation de graphiques et d'histogrammes (cf. Figure 2-10)
- l'usage d'une sémiologie graphique appropriée ; celle-ci remplace l'inventaire classique des formules graphiques par une analyse des moyens et des buts et par un ensemble de règles impératives qui commandent la rédaction graphique, c'est-à-dire le choix des correspondances entre les sensibilités visuelles disponibles et les éléments de l'information. (Bertin, 2005).

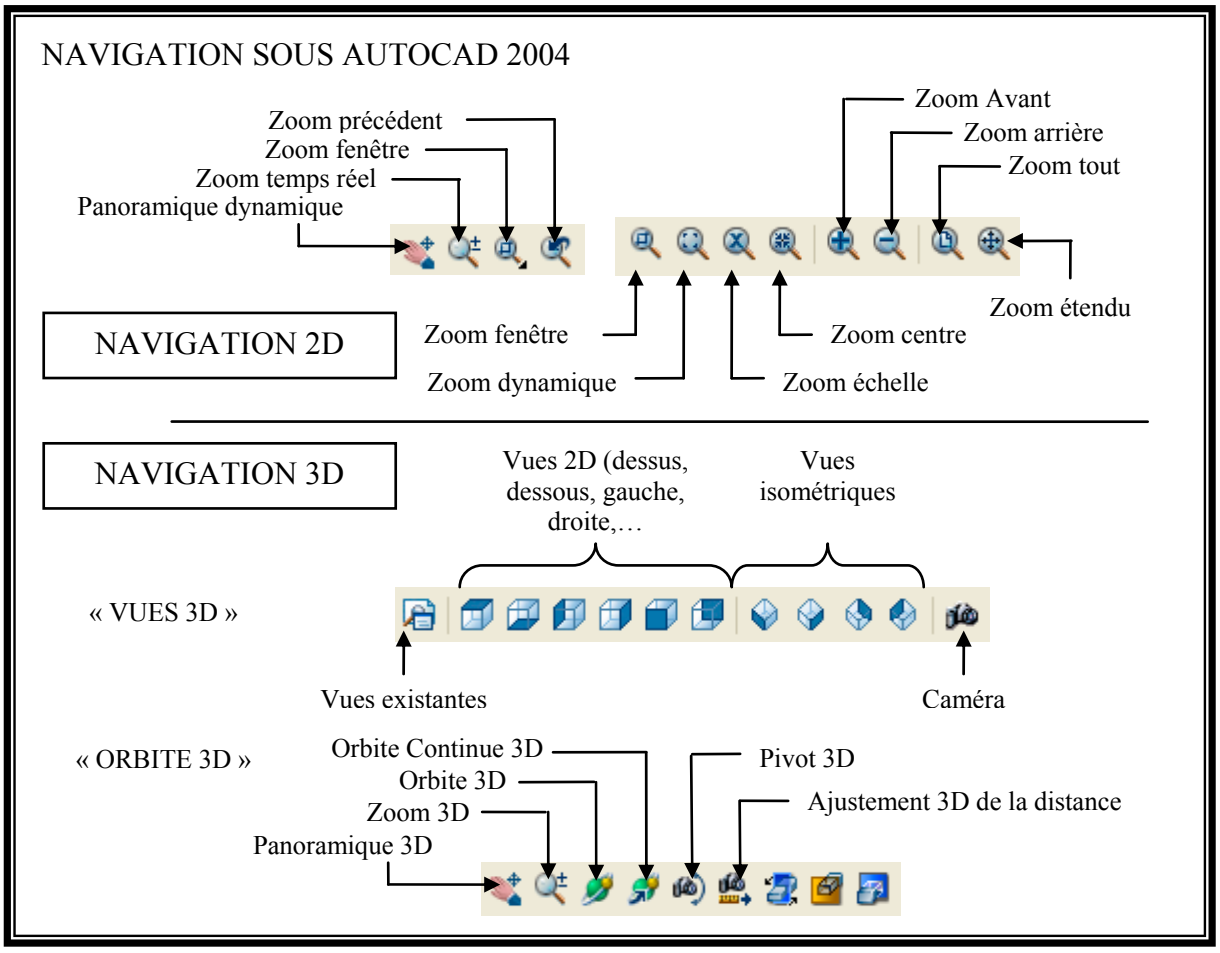


Figure 2-9 : Navigation : zoom et déplacement de l'écran - exemple d'Autocad 2004

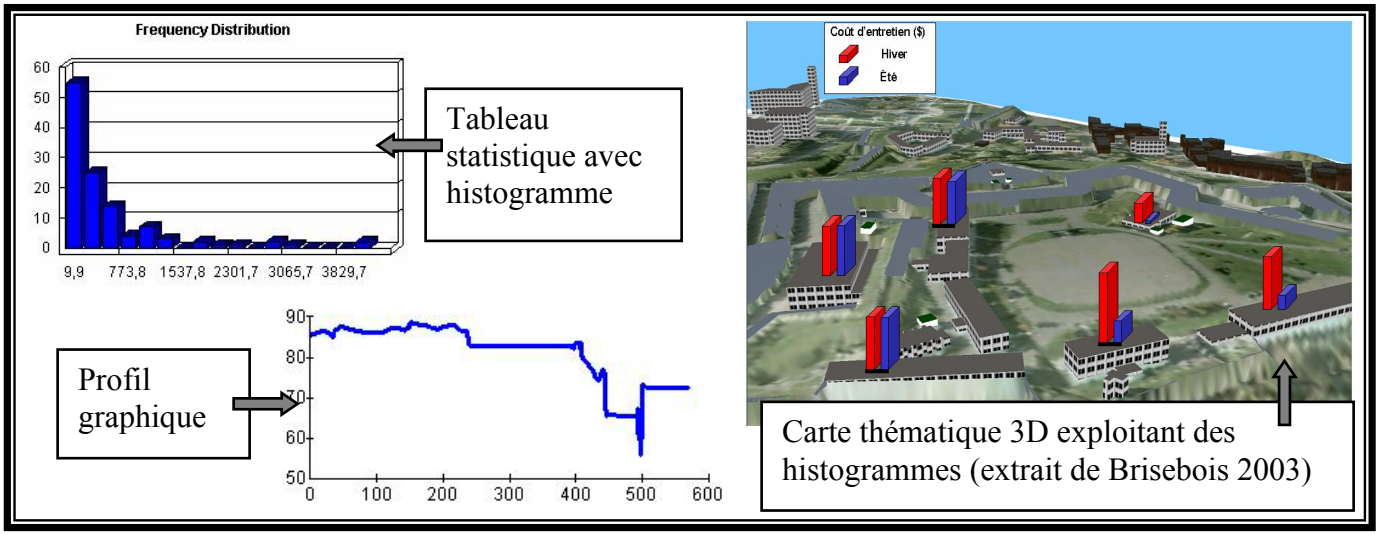


Figure 2-10 : Exemple d'utilisations de graphiques et d'histogrammes

Cette deuxième partie a pu montrer les différentes notions de l'analyse de données et des opérateurs utilisables pour effectuer ces analyses. Notre recherche bibliographique a permis de mettre en avant les principales notions afin de pouvoir optimiser et rechercher de nouveaux concepts d'analyse de données 3D. De surcroît, nous nous sommes intéressés principalement aux analyses spatiales et temporelles (plus spécifiquement aux analyses topologiques) puisque ces analyses intéressent particulièrement les archéologues avec l'analyse stratigraphique.

2.4-Les outils d'exploitation des données spatiales

On entend par le terme « exploitation de données », une exécution des instructions d'un programme par l'unité centrale, qui se traduit en une série d'opérations logiques ou d'opérations de calcul effectuées sur des données ou des informations, entre le moment où celles-ci sont entrées dans un système informatique et celui où elles en sortent (OQLF, 2006). Le terme « traitement de données » (ou traitement des données) est également employé pour désigner le déroulement logique de ces opérations (collecte, enregistrement, fusion, tri, recherche, affichage, modification, impression, etc.), selon un processus prédéterminé (OQLF, 2006). L'exploitation des données passe donc par la gestion (ajout, suppression, modification), l'affichage et l'analyse (descriptive, spatiale, temporelle et

visuelle). Les développements informatiques montrent qu'il existe deux grandes catégories de systèmes : les systèmes transactionnels et les systèmes analytiques.

2.4-1. Les systèmes transactionnels

Les systèmes transactionnels sont des systèmes informatiques dont le fonctionnement repose essentiellement sur le traitement transactionnel, c'est-à-dire sur le traitement immédiat des données (ajout modification suppression), suite à une demande ou à une intervention de l'utilisateur, chaque transaction étant vue comme un ensemble de sous-opérations qui forme un tout fini, avec un début et une fin. La différence entre les concepts de « traitement transactionnel », « traitement transactionnel en ligne » (« OLTP ») et « traitement en temps réel » est si mince, que, la plupart du temps, on trouve les termes « système transactionnel » et « système OLTP » (« On-line Transaction Processing System » (OQLF, 2006).

L'un des systèmes transactionnels les plus connus dans le monde de la géomatique est le **Système d'Information Géographique (ou SIG)**. En effet, le SIG est couramment utilisé pour analyser les données, et les récents développements dans ce domaine sont trop volumineux pour les énumérer (Longley et al., 2005; Nyerges, 2006), notamment puisque les domaines d'applications, dont l'archéologie fait partie (Wheatley et Gilling, 2002), sont riches et vastes. De plus, l'indication du nombre de logiciels[3] et des ouvrages sur le sujet[4] en est une aussi grande preuve. Les SIG sont un outil performant pour : acquérir, gérer, traiter, analyser et diffuser des données géographiques (Bédard, 2003a; Nyerges, 2006). En effet, afin de permettre au système transactionnel d'effectuer l'intégralité des transactions (ajout, modification, suppression) de façon facile, rapide et sûre, la plupart de ces systèmes sont implantés selon une structure relationnelle normalisée de telle sorte que la duplication des données soit à son minimum.

La littérature a montré que les systèmes transactionnels de type SIG performent mieux pour la manipulation de données que pour l'interrogation spatiale (Caron, 1998; Bédard, 2005;

[3] Plus de 70 répertoriés par Geoplan (http://www.geoplan.ufl.edu) en 2001
[4] Sur books.google.com la recherche « GIS » offre plus d'une dizaine de pages d'ouvrages ayant le mot « GIS » dans son titre.

Nyerges, 2006). En effet, les SIG ont aussi une faiblesse lorsqu'il s'agit de traiter un gros volume de données ce qui d'après Caron arrive lorsque les données méritent d'être analysées (Caron, 1998). De plus, pour éviter la duplication des données, normaliser la structure relationnelle revient souvent à augmenter le nombre de tables. Ainsi un nombre important de tables (et donc de jointures entre les tables) et un gros volume de données sont souvent les raisons de la faiblesse des systèmes transactionnels pour analyser les données. En effet, principalement pour des requêtes de types agrégatives et pour celles nécessitant un nombre élevé de jointures entre les tables, une lenteur dans les réponses est toujours constatée lorsque vient le temps d'interroger les données (qui se chiffre en jours plutôt qu'en minutes) (Bédard, 2005; Nyerges, 2006). Ces mêmes requêtes complexifient aussi l'élaboration des requêtes (Bédard, 2005). En effet, même si les dernières années ont vu plusieurs travaux de recherche visant à optimiser et à accélérer l'interrogation, à partir d'un SIG, des bases de données, en utilisant par exemple les requêtes graphiques (Aufore-Portier, 1995; Favetta *et al.*, 2000), une majorité d'outils SIG nécessite encore que l'utilisateur maîtrise tant le langage d'interrogation comme SQL (Egenhoffer *et al.*, 1993) que la structure de la base de données (Langran, 1992).

Bien que les SIG aient d'abord été construits pour gérer des données 2D, ils restent cependant performants quant à la représentation tridimensionnelle des données (Brisebois, 2003). Des effets de perspective et d'ombrage à l'utilisation possible de lunette stéréoscopique en passant par une possibilité d'animation 3D, rendent les SIG comme un outil de visualisation 3D opérationnel. Mais, bien que les SIG offrent une analyse visuelle performante (Nyerges, 2006) et une analyse spatiale bien appropriée au 2D, les SIG commerciaux actuels, aptes à manipuler les données 3D (par exemple : ArcGis - www.Esri.com, MapInfo – www.mapinfo.com, Geomedia – www.intergraph.com), sont plutôt limités à l'analyse des surfaces et à la visualisation 3D. De plus, toutes les analyses se font en 2.5D et non dans un véritable environnement 3D (Tet Kuan *et al.*, 2005) car les SIG, bien qu'ils permettent d'accomplir plusieurs tâches importantes offrent très peu de possibilités de représentation 3D principalement en raison d'une difficulté à structurer les données 3D tant spatiales que descriptives (Lachance, 2005) et à leur incapacité à gérer la plupart des primitives géométriques non planaires (MNT non compris) (Brisebois, 2003).

2.4-2. Les systèmes d'analyse

Les outils d'exploitation de données spatiales devraient servir de support au processus de décision des organisations et permettre de gérer l'historique des données, fournir des réponses rapides et des interfaces à l'usager faciles à utiliser (Bédard, 2005). Les outils qui exploitent les entrepôts de données, le marché de données, les requêteurs et rapporteurs, le forage automatique de données spatiales et les outils basés sur l'approche multidimensionnelle sont des exemples de systèmes d'analyse. L'outil d'analyse, à la manière d'un processus cognitif, doit présenter les données dans une forme proche de la façon dont l'analyste conçoit mentalement le phénomène à analyser. Le travail de déduction qu'il doit effectuer est moins laborieux et son analyse est facilitée (Caron, 1998). De plus, d'après Caron (1998), la **rapidité** et la **facilité** sont les deux conditions que doit rencontrer un système d'analyse pour ne pas nuire au processus cognitif d'analyse de l'usager tel que définit par Champoux (1991). La rapidité permet à l'usager de garder son attention dans sa recherche et son interrogation des données et son processus cognitif (cf. Figure 2-7) n'est pas interrompu. La facilité permet à l'usager de se concentrer sur l'analyse en cours et non sur la manière de formuler son interrogation.

La structure multidimensionnelle : l'analyse rapide et facile des données

Parce qu'elle offre une perspective d'évolution intéressante et qu'une première tentative d'application à l'archéologie a été réalisée (Rageul, 2004), **l'approche multidimensionnelle** sera examinée. À l'intérieur d'une base de données destinée à l'analyse, la dénormalisation des données est utilisée afin de limiter le nombre de tables et le nombre de jointures nécessaires entre les différentes tables dans le but d'améliorer la rapidité d'exécution des requêtes. A la différence des structures transactionnelles qui structurent les données sous la forme de listes unidimensionnelles, la structure multidimensionnelle gère les données sous forme de matrices multidimensionnelles où chaque axe représente un phénomène différent (Caron, 1998) (cf. Figure 2-11).

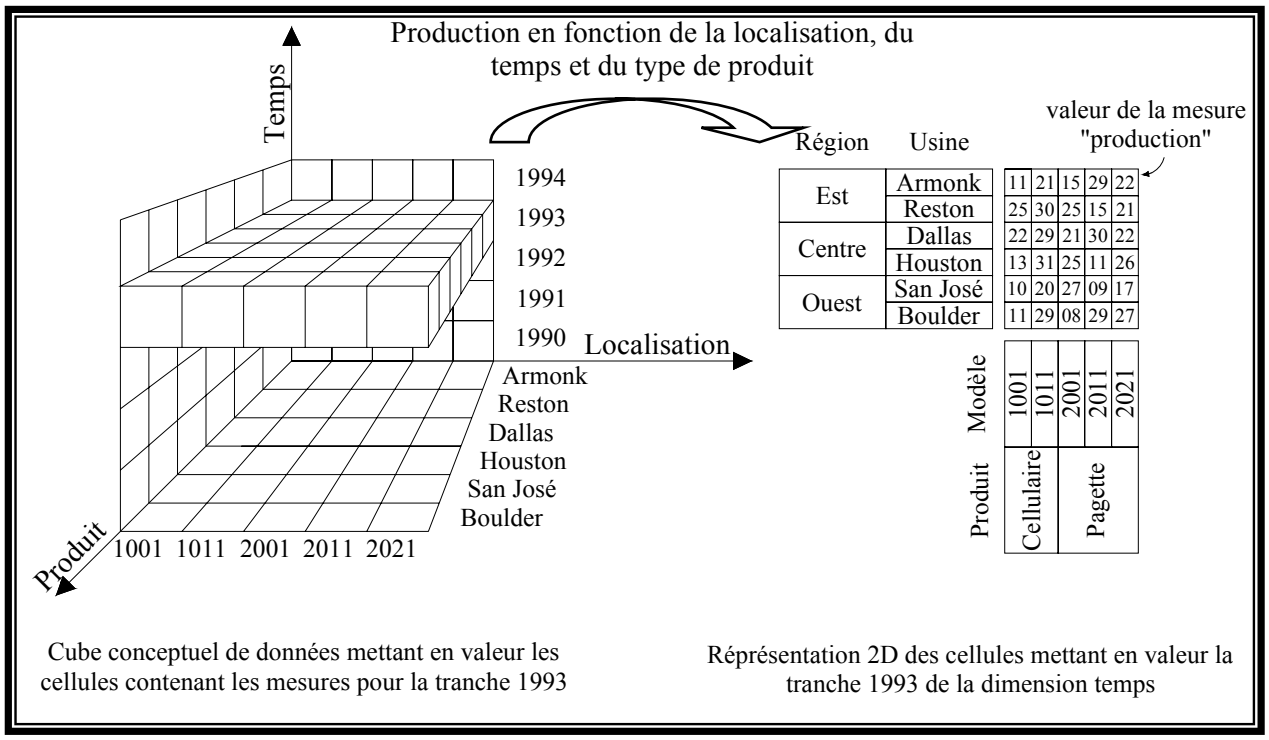


Figure 2-11 : Exemple de cubes, dimensions et mesures inspiré de (Caron, 1998; Bourgon *et al.*, 1999)

Les thèmes d'analyses, ou **dimensions,** sont définis comme les « axes d'analyse selon lesquels on veut étudier des données observables qui, soumises à une analyse multidimensionnelle, donnent aux utilisateurs des renseignements nécessaires à la prise de décision » (OQLF, 2006). D'après la Figure 2-11, la « localisation », le « temps » ou le « produit » sont des exemples de dimensions. Ces dimensions peuvent être hiérarchisées à plusieurs niveaux d'agrégations. (Niveau 1 : Région [Est; Centre; Ouest] - Niveau 2 : Usine [Armonk, Dallas,,…] pour la dimension *Localisation* ou Niveau 1 : Produit [Cellulaire, Pagette] – Niveau 2 : Modèle [1001, 1011,…] pour la dimension *Produit*). À la différence des dimensions où les données peuvent être agrégées, les **mesures** sont les variables non-agrégeables qui font l'objet de l'analyse (dans l'exemple choisi, « production » est une mesure). Elles permettent d'assurer l'unicité de chaque enregistrement (ligne). Finalement, le **Fait** représente la valeur d'une mesure. Il correspond à l'association entre les mesures et les combinaisons de un ou plusieurs membres d'une ou plusieurs dimensions. Dans notre exemple, un fait serait : « L'usine de Dallas de la région du centre a produit 22 cellulaires du modèle 1001 durant l'année 1993 ».

Les **outils OLAP** ou On-Line-Analytical Processing sont définis comme une « catégorie de logiciels axés sur l'exploration et l'analyse rapide des données selon une approche multidimensionnelle à plusieurs niveaux d'agrégation » (Caron, 1998). L'avantage de tels outils renvoi directement aux deux conditions que doit rencontrer un système d'analyse pour ne pas nuire au processus cognitif d'analyse de l'usager ; la facilité et la rapidité. Comme l'analyse se fait à l'aide des dimensions, l'usager n'a plus à maîtriser les langages d'interrogation. Il peut alors interroger les données de manière plus intuitive contrairement aux interfaces souvent complexes des autres systèmes. Quant à la rapidité, les outils OLAP exploite une dénormalisation maximale des données, sous la forme de pré-agrégation stockée. L'usager peut alors obtenir ses réponses dans la seconde même lorsqu'il s'agit d'interrogation portant sur les agrégations et pas seulement sur les données détaillées. Concernant l'analyse de données, il n'y a pas de différence entre un débutant, un expérimentaliste ou un analyste car le principe très intuitif d'outils d'analyse en ligne permet à un n'importe quel utilisateur de devenir opérationnel rapidement.

Les outils OLAP utilisent des opérateurs visuels particuliers afin de « naviguer » dans les cubes multidimensionnels (Rivest, 2000) :

- *Pivoter (pivot, swap)* : Permet d'interchanger deux dimensions. Cette opération est nécessaire comme toutes les dimensions ne peuvent être visualisées en même temps
- *Forer (drill-down)* : Permet de descendre dans la hiérarchie de la dimension, comme par exemple passer d'une analyse des usines du « Centre » à l'analyse des usines de « Dallas »
- *Remonter (drill-up, roll-up)* : Permet de remonter dans la hiérarchie de la dimension et passer d'un niveau détaillé à un niveau plus agrégé, comme par exemple passer d'une analyse des modèles « 1011 » à l'analyse du produit « Cellulaire »
- *Forer latéralement (drill-across)* : pour soit passer d'une mesure à l'autre comme par exemple, visualiser le nombre de personnes travaillant le produit par site au lieu du nombre de produits par sites ou soit passer d'un membre de dimension à un autre comme par exemple, visualiser les données de « Dallas » au lieu de celles de « San José »

Une base de données multidimensionnelle à référence spatiale intègre une ou des classes d'objets spatiaux sous la forme de dimensions spatiales (les membres hiérarchiques de la dimension peuvent avoir ou non une géométrie associée) ou de mesures spatiales dont l'objectif est de pointer vers les objets spatiaux. (Brisebois, 2003). L'avantage d'utiliser une approche multidimensionnelle spatiale avec plusieurs dimensions spatiales, permet à l'utilisateur d'user d'opérateurs visuels, non seulement sur des tableaux ou des graphiques comme le proposerait les outils OLAP, mais aussi sur des occurrences cartographiques, principalement si la dimension spatiale est une dimension spatiale géométrique (i.e. tous les membres de chaque niveau possède une primitive géométrique). Tels sont les concepts des **systèmes SOLAP** ou Spatial OLAP qui permettent alors d'exploiter la composante spatiale des données. Bédard définit ainsi l'outil SOLAP comme « une plate-forme visuelle supportant l'exploration et l'analyse spatio-temporelle faciles et rapides des données selon une approche multidimensionnelle à plusieurs niveaux d'agrégation via un affichage cartographique, tabulaire ou en diagramme statistique » (Bédard, 2005).

La structure multidimensionnelle : la faiblesse pour la gestion des données en temps réel et l'exploitation de la 3ème dimension

Un **système d'analyse temps réel** doit non seulement offrir une compatibilité cognitive avec les utilisateurs potentiels (performance du système : accès rapide et facile à l'interrogation de données), mais il doit aussi leur permettre de pouvoir profiter d'un accès à l'information courante en tout temps (White, 2002). A ce titre, certaines organismes nécessitent un outil d'analyse avec une actualisation plus importante que celle proposée par ce dernier qui se résume généralement à une actualisation périodique (mensuelle, hebdomadaire et rarement quotidienne) et non sur demande. Traditionnellement, il existe aucune connexion directe entre le système d'acquisition et le système d'analyse (Chountas *et al.*, 2004). Cependant, les travaux de Lambert (Lambert, 2005) et White (White, 2002), ont permis de justifier la pertinence d'un outil d'analyse en ligne en temps réel. Cela signifie qu'avec une structure multidimensionnelle en temps réel, le cube ramasserait les données issues de la transaction de façon instantanée (synchronisation entre le système d'acquisition et le système d'analyse concernant l'ajout des données), au lieu d'user d'un processus explicite [(Lachey, 2005) : p29 et p 43]. Cependant, en pratique, l'actualisation des données dans une structure multidimensionnelle correspond juste à l'ajout de données

avant une phase d'analyse et non une gestion entière des données (modification ou suppression).

Concernant **l'exploitation de la troisième dimension**, notre revue exhaustive de littérature indique qu'il ne semble y avoir eu que les travaux de Brisebois relatifs au SOLAP 3D (Brisebois 2003). Tout d'abord, les caractéristiques de l'outil SOLAP 3D sont composées d'une extension des caractéristiques d'outil SOLAP 2D :

1. un traitement de tous les types de primitives géométriques,
2. une analyse spatiale possible.

Les caractéristiques de l'outil SOLAP 3D a aussi des caractéristiques qui lui sont propres :

3. une présence de vues 3D,
4. une navigation tridimensionnelle au sein des vues 3D,
5. une navigation assistée pour explorer le résultat d'une requête, comme le positionnement automatique de l'usager sur le résultat de la requête,
6. une aide à l'orientation,
7. une glissière altimétrique,
8. une source d'éclairage,
9. une fenêtre de manipulation d'objet de petite taille,
10. une génération de vues 2D à partir d'une vue 3D (« slicer »),
11. une possibilité d'impression d'immersion

Cependant, le prototype proposé par Brisebois, présente deux volets : un volet descriptif et un volet cartographique. Ce dernier exploite la composante spatiale géométrique des dimensions et des mesures grâce à des outils de visualisation cartographique de SIG. Or, comme nous l'avons vu à la section 2.4-1. , les SIG sont encore assez limités quant à la gestion explicite de la 3^{e} dimension même s'ils permettent la visualisation dans un univers 3D.

2.5-Conclusion du chapitre

Ce chapitre a pu couvrir l'ensemble des notions que nous avons jugées importantes et pertinentes pour comprendre, optimiser et expérimenter l'analyse de données qui plus est 3D. Le contexte archéologique pose plusieurs défis à la géomatique.

Le premier défi concerne l'aspect tridimensionnel des données. Face à une diversité dans la notion de 3D, nous avons associé, pour le présent mémoire, le terme d'objet 3D à tout objet à trois dimensions géométriques (longueur, largeur et hauteur) et positionné dans un univers 3D (x, y, z) et référencés temporellement. La première partie de ce chapitre a aussi mis en avant le rôle prépondérant de la donnée dans la représentation numérique de cet objet 3D.

Le deuxième défi que le contexte archéologique lance à la géomatique concerne la nécessité d'analyser simultanément l'espace et le temps, dans un contexte où l'un peut être influencé par l'autre, et de le faire rapidement et facilement dans un environnement tridimensionnel afin de mieux supporter l'interprétation des données. Ainsi, la deuxième partie de ce chapitre a apporté les principales notions théoriques et pratiques concernant l'analyse des données. De surcroît, nous nous sommes intéressés principalement aux analyses spatiales et temporelles (plus spécifiquement aux analyses topologiques) puisque ces analyses intéressent particulièrement les archéologues avec l'analyse stratigraphique. La troisième partie de ce chapitre a aussi permis de mettre en évidence deux approches globales pour exploiter les données spatiales. Ces deux approches se distinguent essentiellement lorsqu'il s'agit de traiter un gros volume de données tridimensionnelles. **Les systèmes transactionnels** sont des systèmes informatiques dont le fonctionnement repose essentiellement sur le traitement immédiat des données (ajout, modification et suppression) au détriment d'une analyse rapide et facile. **Les systèmes analytiques**, quant à eux, à la manière d'un processus cognitif, doivent présenter les données dans une forme proche de la façon dont l'analyste conçoit mentalement le phénomène à analyser.