CONCEPTION ET ÉVALUATION D'UN ENTREPÔT DE DONNÉES HÉTÉROGÈNES POUR LE SUIVI DU RÔLE DE L'AGRICULTURE SUR LE TERRITOIRE DU LANGUEDOC-ROUSSILLON

DEFINITION, DESIGN AND EVALUATION OF A DECISION SUPPORT SYSTEM BASED ON HETEROGENEOUS DATA FOR AGRICULTURE MONITORING IN THE LANGUEDOC-ROUSSILLON TERRITORY

Établissement **institut des sciences et industries du vivant et de l'environnement (AgroParisTech)**
École doctorale **GAIA - Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau**
Spécialité **STE - Sciences de la Terre et de l'Eau**
Unité de recherche **TETIS - Télédétection Environnement Télédetection et Information Spatiale**
Directeur de la thèse Flavie CERNESSON
Co-Directeur Maguelonne TEISSEIRE
Co-Encadrant Lucile SAUTOT

Concours GAIA
Début de la thèse le **1 octobre 2018**
Date limite de candidature **16 mai 2018**

Description de la problématique de recherche

Le diagnostic et le suivi en ingénierie territoriale s'appuient sur de nombreuses données hétérogènes issues de sources variées : images satellites, cartes d'artificialisation des sols, modèles numériques de terrain, données agricoles, données environnementales, documents de planification et réglementaires etc.

L'hétérogénéité des données est une notion qui intègre plusieurs composantes : 1) la nature des données (tableaux de mesures

qualitatives, tableaux de mesures quantitatives, tableaux de mesures mixtes, images, textes, graphes) ; 2) les échelles spatiotemporelles auxquelles peuvent être récoltées les données ; 3) les thématiques auxquelles se rattachent les données. Ces différentes composantes sont indissociables et complexifient la mise en relation de données dans les systèmes d'information.

Les outils SIG classiquement utilisés pour l'exploration de données spatiales permettent ainsi de faire le lien entre une thématique et objets spatiaux au travers d'une couche de données spatiales associée à une table attributaire. Les SIG permettent d'agréger les informations sur une couche, et les opérations d'analyse sont menées pas à pas en manipulant une ou deux couches à la fois. Les entrepôts de données sont l'un des types phares de plate-formes numériques permettant le stockage, le croisement, l'agrégation et l'analyse de gros volumes de données. Cependant, la littérature scientifique montre que cette question est loin d'être réglée, que ce soit du point de vue théorique ou du point de vue opérationnel. En effet, la conception des entrepôts reste rigide en considérant les « faits » devant être quantitatifs et les « dimensions » devant être des catégories organisées en hiérarchie. « Faits » et « dimensions » dépendent donc de la nature des données. De nombreux travaux existent sur les composantes « échelles spatio-temporelles » et « thématiques ».

Les travaux qui se sont attachés à l'intégration de données de nature différente (images, textes, graphes, etc.) au sein des entrepôts de données à la base des systèmes d'information décisionnels [Lewis et al, 2017, Zhao at al., 2011] sont moins nombreux. Ces travaux ont proposé différents modèles, adaptés à chaque type de données [Lewis et al, 2017, Kasprzyk, 2015 ; Mendoza et al., 2015 ; Sautot et al., 2014]. Cependant, peu de travaux se sont penchés sur (1) l'analyse de contenu des textes comme support aux dimensions d'analyse [Bringay et al., 2011] et (2) sur le croisement de données de différentes natures [Boukraâ at al., 2010; Loudcher et al., 2015]], par exemple le croisement d'images et de textes au sein d'un entrepôt de données. Tous ces travaux restent partiels compte tenu de la nature des données que nous aurons à traiter.

La thèse comporte donc un premier verrou qui consiste, tout en s'appuyant sur les savoirs existant, en la proposition d'un nouveau paradigme de modélisation des données, afin de pouvoir réaliser des agrégations et des analyses sur des données de natures différentes.

Un second verrou porte sur l'évaluation des performances de ce type d'outil, qui, ne peut passer qu'au travers de cas concret impliquant des utilisateurs pouvant être non-informaticiens.

Ce sont à ces verrous que les travaux proposés dans cette thèse souhaitent répondre en (1) définissant un modèle d'entrepôt de données permettant la gestion et le croisement de données hétérogènes et (2) offrant les outils d'analyse de ces nouveaux types d'information au travers d'un cas concret mettant en avant le lien agriculture/territoire en Languedoc-Roussillon.

Le rôle de l'agriculture sur un territoire est majeur car bien que peu nombreux, les agriculteurs gèrent de grandes étendues de l'espace. Par exemple pour l'Hérault, les agriculteurs représentent 6% de la population et les zones agricoles 30% de la surface du département. En Languedoc-Roussillon, assurer la pérennité des exploitations dans un contexte très concurrentiel du point de vue notamment rdesessources sol et eau, ainsi qu'une production agricole répondant à des exigences de qualité et respectueuse d'un milieu particulièrement vulnérable concernent tous les acteurs du territoire (gestionnaire du territoire, agriculteurs…). Les progrès réalisés ces dernières années en informatique étendent le panel des informations disponibles pour les acteurs et facilitent l'accès aux outils d'aide au diagnostic, au suivi voire à la prise de décision. Travailler sur un cas concret défini avec les acteurs de terrain renforce la pertinence de la question traitée tant sur le plan informatique qu'en sciences de l'environnement.

La thèse s'appuiera sur 4 grandes étapes de travail :

1) Collecte, appropriation des données et définition des indicateurs d'usage

2) Modélisation conceptuelle de l'entrepôt

3) Modélisation logique et optimisation des structures de données (agrégation et mise en lien)

4) Evaluation sur un cas d'étude concret.

Elle s'appuiera sur des collaborations fructueuses tant académiques (avec le LISAH et HSM) qu'avec les acteurs de terrain (DRAF Occitanie, DDTM 34 notamment).

Objectifs

Les entrepôts de données facilitent l'exploration d'un grand volume de données en pré-calculant un ensemble d'indicateurs, jugés pertinents par les utilisateurs identifiés du système, et en offrant la possibilité aux-dits utilisateurs de consulter ses indicateurs à différentes échelles spatio-temporelles.

L'objectif visé est de (1) concevoir des outils informatiques facilitant le croisement et l'agrégation des données de différents types (mesures, documents, images etc.), au sein d'un unique entrepôt (2) évaluer les performances de l'entrepôt sur un cas pratique de diagnostic territorial focalisé sur le rôle de l'agriculture dans des démarches de développement territorial en Languedoc-Roussillon.

Project description

Diagnosis and monitoring in territorial engineering are based on numerous heterogeneous data from various sources: satellite images, artificial soil maps, digital terrain models, agricultural data, environmental data, planning and regulatory documents, etc.

Data heterogeneity is a concept that integrates several components: 1) the nature of the data (qualitative measurement tables, quantitative measurement tables, mixed measurement tables, images, texts, graphs); 2) the spatial and temporal scales at which the data can be collected; 3) the themes to which the data relate. These different components are inseparable and complicate the linking of data in information systems.

The GIS tools conventionally used for spatial data exploration thus make it possible to make the link between a theme and spatial objects through a spatial data layer associated with an attribute table. GIS allows information to be aggregated on a layer, and analysis operations are carried out step by step by manipulating one or two layers at a time. Data warehouses are one of the leading types of digital platforms for storing, crossing, aggregating and analyzing large volumes of data. However, the scientific literature shows that this issue is far from being resolved, either from a theoretical or an operational point of view. Indeed, the design of warehouses remains rigid by considering the'facts' that should be quantitative and the'dimensions' that should be categories organized in a hierarchy. 'Facts' and'dimensions' therefore depend on the nature of the data. Many works exist on the'space-time scales' and'thematic' components. Less work has been done to integrate data of a different nature (images, texts, graphs, etc.) into data warehouses that form the basis of decision-making information systems (Lewis et al., 2017, Zhao at al., 2011). These studies proposed different models, adapted to each type of data[Lewis et al, 2017, Kasprzyk, 2015; Mendoza et al., 2015; Sautot et al., 2014]. However, little work has been done on (1) content analysis of texts as a support to analytical dimensions[Bringay et al., 2011] and (2) cross-referencing of data of different natures[Boukraâ at al., 2010; Loudcher et al., 2015], for example crossing images and texts within a data warehouse. All this work remains partial given the nature of the data we will have to process.

The thesis thus includes a first lock which consists, while relying on existing knowledge, in the proposal of a new paradigm of data modeling, in order to be able to carry out aggregations and analyses on data of different natures. A second lock concerns the evaluation of the performance of this type of tool, which can only pass through concrete cases involving users who may not be computer scientists.

It is to these obstacles that the work proposed in this thesis wishes to answer by (1) defining a data warehouse model allowing the management and the crossing of heterogeneous data and (2) offering the tools of analysis of these new types of information through a concrete case highlighting the agriculture/territory link in Languedoc-Roussillon.

The role of agriculture in a territory is major because although few in number, farmers manage large tracts of land. For example, for the Hérault, farmers represent 6% of the population and agricultural areas 30% of the surface area of the department. In Languedoc-Roussillon, ensuring the sustainability of farms in a highly competitive context, particularly from the point of view of soil and water resources, as well as agricultural production that meets quality requirements and respects a particularly vulnerable environment concerns all stakeholders in the territory (land managers, farmers, etc.). The progress made in recent years in information technology has extended the range of information available to stakeholders and facilitated access to tools for diagnosis, monitoring and even decision-making.

Working on a specific case defined with the actors in the field reinforces the relevance of the issue dealt with both in terms of information technology and environmental sciences.

The thesis will be based on 4 main work steps:

1) Collection, appropriation of data and definition of use indicators

2) Conceptual modelling of the warehouse

3) Logical modelling and optimization of data structures (aggregation and linking)

4) Evaluation on a concrete case study.

It will be based on fruitful collaborations, both academic and with actors in the field.

Mots-clés

Territoire,
Transition numérique,
Entrepôts de données,
Développement durable,
Données hétérogènes,
Languedoc Roussillon

Keywords

Territoire,
Digital transition,
Data warehouse,
Sustainable development,
Heterogeneous data,
Languedoc Roussillon

Profil et compétences recherchées

M2 recherche ou ingénieur en informatique avec un fort intérêt pour les sciences de l'environnement ou l'agronomie ou en sciences de l'environnement/agronomie avec une spécialisation en informatique (type Agrotique ou hydro-informatique) ou.
Des compétences avancées en programmation (Java, C++ ou python) et en base de données sont requises.

Profile and skills required

Master degree in Computer Sciences with a great interest for environmental sciences or agronomy, or Environmental Sciences /Agronomy with a specialized in computer sciences (Agro-informatics, Hydro-informatics).
Avanced skills in programming (Java, C++ ou python) and databases.

Localisation

La thèse aura lieu à Montpellier :
Maison de la Télédétection
500 rue Breton
34000 Montpellier
FRANCE

The thesis is located at:
Maison de la Télédétection
500 rue Breton
34000 Montpellier
FRANCE

Contacts :

Lucile Sautot :
lucile.sautot@agroparistech.fr
+33 4 67 55 86 19

Maguelonne Teisseire
maguelonne.teisseire@irstea.fr

Flavie Cernesson
flavie.cernesson@agroparistech.fr