STedi : une infrastructure logicielle pour renforcer la qualité des données territoriales statistiques

Camille Bernard, Benoit Le Rubrus, Marlène Villanova-Oliver, Jérôme Gensel, Ronan Ysebaert, Isabelle Salmon

STedi : une infrastructure logicielle pour renforcer la qualité des données territoriales statistiques

Camille Bernard[1], Benoit Le Rubrus[1], Marlène Villanova-Oliver[1], Jérôme Gensel[1], Ronan Ysebaert[2], Isabelle Salmon[2]

1. LIG – Equipe STeamer, rue de la Passerelle, 38402 Saint Martin d'Hères, France
2. UMS RIATE, Université Paris 7 - UFR GHSS, 75205 Paris, France

RESUME. *L'information à référence spatiale doit aujourd'hui respecter des normes pour garantir l'interopérabilité des systèmes de données. Si les normes ISO recommandées par la directive INSPIRE en vigueur sont adaptées aux données environnementales, mettre en œuvre la directive pour des données statistiques territoriales nécessite des adaptations. Le modèle de données INSPIRE est alors étendu pour la représentation de ces données, dans le cadre du projet « ESPON Database 2013 ». Cet article présente l'infrastructure logicielle « STedi » et ses composants reposant sur le modèle INSPIRE étendu proposé. Cette infrastructure opérationnelle permet la gestion d'un flux de données dans son ensemble : du fournisseur à l'utilisateur, de l'acquisition à la restitution des données et métadonnées. L'infrastructure veille au contrôle des données et métadonnées recueillies par le biais de briques logicielles telles que le « Checking tool », chargée d'organiser le processus de vérification de la qualité des jeux de données. Le caractère modulaire et le respect de standards font de l'infrastructure un système transposable à de multiples domaines.*

ABSTRACT. *Spatial information must now meet standards to ensure data systems interoperability. While ISO standards, recommended by the INSPIRE directive in force, are adapted to environmental data, still many adjustments are necessary to adapt standards to territorial statistical data. Therefore, the INSPIRE data model is extended as part of the "ESPON Database 2013" project. This article presents the "STedi" software infrastructure and its components based on the INSPIRE extended model proposed. This operational infrastructure enables the management of a data flow as a whole: from the provider to the user, from the acquisition to the retrieval of data and metadata. The infrastructure provides the oversight of data and metadata collected through software components such as the "Checking tool", which is in charge of datasets quality process organisation. The modular infrastructure is compliant with standards and is transposable to multiple domains.*

MOTS-CLES : *INSPIRE, modèle de données, qualité des données, infrastructure logicielle.*

KEYWORDS: *INSPIRE, data model, data quality, software infrastructure.*

1. Introduction

En matière de production d'informations statistiques, les fournisseurs sont incités à diffuser conjointement les données et leurs métadonnées. En effet, connaître la source des données ou la description de la méthodologie employée pour la construction d'un indicateur permet aux utilisateurs d'appréhender et d'interpréter correctement les phénomènes statistiques décrits, en particulier dans le contexte de comparaisons internationales (OCDE, 2002).

Partant du constat de la primordialité des métadonnées, le consortium du projet *Multi-Dimensional Database Design and Development* (M4D) construit, dans le cadre du programme *ESPON 2013*[1], l'infrastructure logicielle *STedi* pour la gestion d'indicateurs statistiques à références spatiale et temporelle. L'enjeu du projet M4D consiste, entre autres, en l'harmonisation et la mise à disposition de jeux de données produits par des fournisseurs multiples. Le modèle créé est conforme à la directive INSPIRE, par obligation légale, pour garantir l'interopérabilité du système. Si le modèle de données INSPIRE est adapté aux données environnementales (IGN, 2012), il est nécessaire de l'étendre pour appréhender les particularités de l'information statistique territoriale (Plumejeaud *et al.*, 2010). Ce modèle étendu doit être strictement respecté par les producteurs pour garantir l'harmonisation, la conformité et la qualité des données recueillies.

A cette fin, l'outil interactif de suivi en ligne *Checking tool*, est créé. Cet outil, de l'infrastructure *STedi*, aide le producteur de données à améliorer la qualité de ses dépôts par le biais de rapports d'expertise. Ces rapports deviennent les documents de référence pour l'évaluation et l'amélioration de la qualité des jeux de données (Bergdahl et al., 2007) avant leur intégration en base de données. A l'autre bout de la chaine, un outil de recherche en ligne permet d'interroger et de consulter les données, pour lesquelles sont générées des fiches de métadonnées précises.

Cet article présente l'infrastructure logicielle *STedi* pour *Spatio Temporal evolutive data infrastructure*, qui organise la gestion de ce flux de données : de la collecte des données et métadonnées à leur restitution, du fournisseur à l'utilisateur. Cette infrastructure logicielle opérationnelle, modulaire et évolutive, établit les liens entre les outils (briques logicielles) et le modèle de données INSPIRE étendu. En veillant à la qualité des données et métadonnées, elle garantit une restitution riche et précise de l'information statistique. Le respect des normes et l'architecture modulaire confèrent au système une capacité d'adaptation à divers domaines.

La section 2 de cet article expose plus précisément le contexte de ce travail. La section 3 rappelle les spécificités du modèle de données implémenté au sein de l'infrastructure *STedi*. La section 4 décrit le composant « *Checking Tool* ». La section 5 présente le système *STedi*, un outil flexible, coordonnant les composants logiciels pour la gestion du flux de données dans son ensemble.

---

[1] *European Observation Network for Territorial Development and Cohesion* http://www.espon.eu

2. Contexte

2.1 Le projet M4D du programme ESPON

Le programme *ESPON 2013*, adopté en novembre 2007 par la Commission Européenne, regroupe une série de projets menés par des Groupes de Projet Transnationaux (TPG) produisant des analyses de l'ensemble du territoire européen. Les projets couvrent des domaines de recherche variés pour soutenir l'élaboration de politiques en matière de développement et de cohésion territoriale. Dans le cadre de ce programme, le consortium M4D est notamment en charge de la création d'une base de données visant, d'une part, à recueillir les données produites par les différents TPG, d'autre part, à les restituer via une interface de requête en ligne, destinée à un public de chercheurs et de politiques.

Dans les phases précédentes du programme, les données restituées par les TPG sont hétérogènes. Le groupe M4D procède alors à une analyse des caractéristiques des données en vue de leur harmonisation.

2.2 Caractéristiques des données statistiques territoriales du programme ESPON

Malgré leur hétérogénéité, les données des projets ESPON ont en commun d'être des données statistiques reposant sur des nomenclatures territoriales. Les jeux de données sont construits à partir de données multiples mobilisées à des fins d'analyse, de prospection, de conseil et d'aide à la décision. Ils concernent des points d'intérêts de la politique de Cohésion de l'Union Européenne.

Dans l'article Plumejeaud *et al.* (2010), nous avons mis en évidence la structure de l'information statistique, généralement présentée sous la forme de jeux de données, regroupant chacun une collection d'indicateurs, mesurés de façon spécifique, pour un ensemble d'unités territoriales, à différentes dates. Le jeu de données statistique peut être appréhendé selon trois niveaux d'information : le niveau du jeu de données, le niveau des indicateurs composant le jeu, et le niveau de la donnée, décrivant les valeurs des indicateurs pour chacune des unités statistiques.

Là où des organismes comme Eurostat compilent des indicateurs de données statistiques récoltés par des Instituts Nationaux, les projets ESPON constituent des jeux de données complexes portant sur l'ensemble du territoire européen, et parfois son voisinage, à différentes échelles, dans différents maillages territoriaux. En outre, ils proposent des indicateurs innovants reposant sur des méthodologies éprouvées scientifiquement (typologies, scénarios). Les jeux de données peuvent combiner plusieurs indicateurs et sources de données. La combinaison de différentes sources de données engendre des indicateurs composites dont les valeurs sont parfois estimées. Les jeux de données reposent sur une ou plusieurs nomenclatures du

territoire, parfois sur plusieurs niveaux (par exemple NUTS[2] 0, 1, 2, 3), et pour différentes versions (1999, 2003, 2006) créées suite à des changements territoriaux[3]. Les objets géographiques traités sont multiples : NUTS et objets urbains tels que les *Urban Mophological Zones* (UMZ), *Functional Urban Area* (FUA), *Morphological Urban Area* (MUA).

L'information statistique délivrée par les projets ESPON est multidimensionnelle et complexe. L'objectif est alors de mettre en place un modèle générique, conforme aux normes pour la diffusion de données géographiques et décrivant suffisamment l'information statistique pour sa compréhension et sa reproductibilité. La section suivante présente le modèle de données élaboré, dont une implémentation répond aux besoins du programme ESPON.

3. Un modèle pour les données et métadonnées statistiques territoriales

3.1 Extension du modèle de métadonnées préconisé par la directive INSPIRE

Le guide d'implémentation des métadonnées INSPIRE, à disposition des diffuseurs de données géographiques, décrit un ensemble de règles auxquelles l'information à référence spatiale doit se conformer (European Commission Joint Research Centre, 2013). Le modèle de données détaillé dans ce guide est basé sur les normes ISO 19115 (norme définissant le schéma requis pour décrire des informations géographiques et des services au moyen de métadonnées[4]) et ISO 19119 (norme définissant des schémas architecturaux relatifs aux interfaces de service utilisées pour les informations géographiques[5]). S'il s'avère que ce modèle est particulièrement adapté à la représentation de données géographiques environnementales (European Commission Joint Research Centre, 2013), il l'est moins à celle de données reposant sur des nomenclatures d'unités territoriales (Ysebaert *et al.*, 2014). Le modèle de données que nous proposons pour le programme ESPON répond à cette limitation. Grâce à nos travaux antérieurs (Plumejeaud *et al.*, 2010), nous avons élaboré un profil de la norme ISO 19115, nommé *esponMD*. La directive INSPIRE permet de décrire une ressource selon un titre, un résumé, des mots clés, des thèmes, etc. Dans le modèle proposé, nous considérons le jeu de données comme une ressource. L'extension permet de spécifier le contenu du jeu de données (indicateur(s) et source(s)) en ajoutant la balise « esponMD:datasetContentInfo » sous la balise ISO 19115 « gmd:contentInfo ». Une seconde extension permet de décrire plus précisément la dimension spatiale du jeu de données par la balise « esponMD:spatialBinding », balise fille de la balise ISO 19115 « gmd:geographicElement ».

---

[2] Nomenclature d'Unités Territoriales Statistiques. Un Etat européen est au niveau NUTS 0.
[3] http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/history_nuts
[4] http://www.iso.org/iso/fr/home/store/catalogue_tc/catalogue_detail.htm?csnumber=53798
[5] http://www.iso.org/iso/fr/home/store/catalogue_tc/catalogue_detail.htm?csnumber=39890

Le profil *esponMD* permet ainsi de décomposer les jeux de données statistiques territoriaux jusqu'à l'entité élémentaire que représente la valeur d'un indicateur pour une unité spatiale et une période (ou date) données.

Le profil permet aussi de spécifier les éléments de métadonnées relatifs à la méthodologie employée pour la construction de l'indicateur (*Methodology*) ou la nature de l'indicateur (attribut *NatureType*) et l'unité de mesure utilisée (*UnitOfMeasure*). Des métadonnées pour la description des sources de données composant le jeu (*SourceReference*) sont reliées directement dans le modèle au niveau de la valeur d'un indicateur (*ValueRegistry*). Ce niveau de précision permet, lors de la restitution des données, de retrouver la source de chaque donnée indépendamment les unes des autres.

L'ensemble des champs de métadonnées rendus obligatoires par la directive INSPIRE est implémenté au sein du modèle de données que nous proposons. La figure 1 présente un extrait des principales entités du modèle, distribuées dans trois schémas : métadonnées, données et spatial.

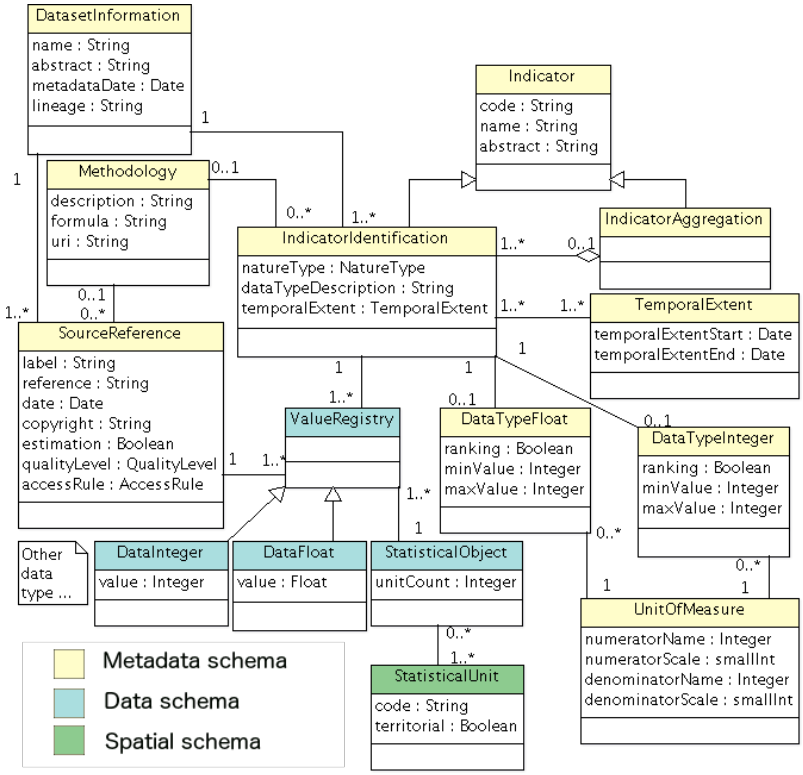

Figure 1 - Extrait du modèle de données ESPON

3.2 Spécificités du modèle

Le modèle INSPIRE étendu par le profil *esponMD* permet de prendre en compte les caractéristiques des données statistiques territoriales combinant plusieurs sources de données et portant sur des objets géographiques emboîtés (divisions territoriales ou administratives). La forme standardisée/normalisée du modèle assure, en aval, la circulation des données vers des systèmes respectant, eux aussi, la directive INSPIRE. Nous présentons dans cette section les spécificités du modèle créé.

3.2.1 Richesse des métadonnées

L'ensemble *Metadata* du modèle de données regroupe des métadonnées à propos : (i) de l'identité d'un jeu de données et de son extension spatiale et temporelle, (ii) des contacts associés au jeu de données, (iii) des indicateurs composant un jeu de données et leurs emprises spatiale et temporelle, (iv) des sources, des méthodologies, et des unités de mesure utilisées pour la constitution des valeurs de l'indicateur, (v) des droits d'utilisation et de distribution des données. Les métadonnées décrivent donc les trois sous-éléments principaux du modèle : le jeu de données, l'indicateur et la valeur.

Eurostat, organisme spécialisé dans la diffusion d'informations statistiques, propose également un modèle pour représenter les spécificités de l'information statistique : *EURO-SDMX*[7] *Metadata Structure* (ESMS). Il supporte des métadonnées relatives à la méthodologie employée pour la construction de l'indicateur (Götzfried, 2009) mais il semble représenter succinctement la dimension spatiale de l'information. Le modèle que nous proposons est particulièrement adapté aux données statistiques (section 3.2.2) et à la représentation des évolutions des divisions du territoire, comme nous le montrons dans la section 3.2.3.

3.2.2 Finesse du modèle

Notre modèle permet de représenter l'information statistique selon des niveaux dépendants les uns des autres : nous retrouvons au plus haut niveau le jeu de données, puis l'indicateur, et enfin la valeur d'un indicateur. Le niveau de granularité du modèle étant la valeur d'un indicateur pour un temps donné et une unité spatiale donnée, il est possible, en sortie, de filtrer l'information pour obtenir la valeur d'un indicateur à ce niveau de granularité.

De plus, la finesse du modèle permet d'associer une source à chacune des valeurs d'un indicateur (cf. figure 1), et non pas uniquement au jeu de données. Il est ainsi possible de déterminer les sources utilisées pour la construction de l'indicateur et de retrouver la source dont est issue une des valeurs de l'indicateur pour une unité spatiale et temporelle particulière. Les métadonnées relatives à la méthodologie permettent de comprendre comment un indicateur a été construit et de rendre compte de la complexité des indicateurs créés par les projets ESPON.

---

[7] *Statistical Data and Metadata eXchange* (SDMX)

3.2.3 Dimension spatiale et temporelle

Les valeurs d'un indicateur peuvent être renseignées pour plusieurs dates et périodes temporelles. Quant à la dimension spatiale, une base de données qui implémente le modèle peut stocker les valeurs d'un indicateur pour différents objets géographiques, sur plusieurs niveaux (par exemple, les NUTS 0, 1, 2, 3) et selon différentes versions de définition des objets (1999, 2003, 2006). Toute nomenclature du territoire composée de un ou plusieurs niveaux géographiques, respectant le format tabulaire attendu, peut donc être intégrée à l'infrastructure *STedi*.

Le modèle supporte plusieurs versions d'une même nomenclature, les relations entre ces versions permettent de tracer l'historique des modifications administratives opérées sur les territoires. La typologie des événements affectant une ou plusieurs unités spatiales entre deux versions de nomenclature est dressée pour qualifier les événements ayant entrainé la modification de ces unités, qu'il s'agisse de fusion, de scission, d'intégration, d'extraction, etc. (Plumejeaud, 2011). Également, la typologie des dérivations éventuelles entre deux unités spatiales de différentes nomenclatures est établie en quatre classes : *includes, equals, included, intersects*. Sur les bases de cette typologie, il est ainsi possible d'envisager des outils automatisant le transfert des valeurs d'une version de nomenclature à une autre. Par exemple, si la base de données contient des valeurs collectées de population définies pour l'année 1970 dans la version 2010 de la nomenclature NUTS, alors ces valeurs pourront être estimées égales pour des unités territoriales inchangées (*equals*) dans des versions de nomenclature ultérieures.

Le modèle de données présenté est exploité par tous les outils implémentés au sein de *STedi*. Notamment, des outils de recherche permettent à l'utilisateur de télécharger l'information (données et métadonnées d'indicateurs) filtrée selon des critères issus des champs de métadonnées prévus par le modèle (mots clés, nature des indicateurs, filtres temporel et/ou spatial). Nous ne décrivons pas ici ces outils d'accès aux données[8]. Nous nous focalisons sur un des composants de *STedi* nommé « *Checking Tool* ». Cet outil permet de centraliser la récolte des données, de vérifier leur conformité avec le modèle. Il contribue ainsi à l'amélioration de la qualité des jeux de données.

4. Le composant « Checking tool »

Nous présentons ici un outil de l'infrastructure *STedi,* opérationnel et en ligne, permettant de regrouper des acteurs autour du recueil de données.

Le consortium M4D doit assurer la qualité des données recueillies. Les critères de qualité sont définis en fonction de trois grands types de jeux de données : les *Key*

[8] Ysebaert et al. (2014) décrit l'outil de restitution des jeux de données du type *Key Indicators*.

Indicators (leurs données couvrent l'ensemble du territoire européen et reposent sur des nomenclatures du territoire) ; les *Case Studies* (leurs données peuvent porter sur des régions européennes ou non) ; les *Background Data* (des ensembles de données et matériels jugés utiles par les projets). Des procédures de vérification différentes sont élaborées pour ces trois types : plus le degré d'exigence concernant la qualité d'un jeu de données est élevé, plus les étapes de vérification sont nombreuses.

Les jeux de données de type *Key Indicators* font l'objet d'un suivi particulier : le niveau d'exigence pour ces derniers, quant à la qualité des données et des métadonnées, est élevé. Nous présentons ici uniquement le suivi de ces jeux de données. Les critères de qualité principaux des jeux de données *Key Indicators* sont les suivants : (i) si l'indicateur est issu d'une analyse statistique, la méthodologie employée doit être suffisamment détaillée pour permettre la reproductibilité du calcul des valeurs ; (ii) les métadonnées doivent préciser les sources et les méthodes employées pour la création des indicateurs ; (iii) les données d'un jeu doivent être disponibles pour un maximum d'unités statistiques de la nomenclature territoriale choisie ; (iv) idéalement, les valeurs manquantes doivent être estimées. La méthodologie employée est alors décrite dans les métadonnées du jeu (Ysebaert, Le Rubrus, 2012). Au delà des critères de qualité, les indicateurs contenus dans un jeu doivent être innovants et non déjà inclus dans la base de données, à moins de mettre à jour un indicateur existant.

Le groupe M4D devait disposer d'un outil souple permettant de paramétrer des étapes de vérification et des catégories d'acteurs intervenant au cours du processus de recueil des jeux de données. C'est pour répondre à ce besoin que l'outil de suivi en ligne *Checking tool* a été créé. Embarqué au sein de l'application web *ESPON Database Portal* (http://database.espon.eu), il permet de collecter et vérifier les jeux de données, et dans le cas des *Key Indicators*, d'en améliorer la qualité, avec l'aide du fournisseur. Le processus est composé de différentes étapes de vérification automatisées ou réalisées par des experts. Les acteurs (producteurs de données, spécialistes des métadonnées, statisticiens et commanditaires du programme) interagissent pour améliorer la qualité des jeux de données, par le biais de commentaires déposés à chaque étape. Ils ont la possibilité de vérifier et modifier les métadonnées, de contrôler et corriger les valeurs extraordinaires identifiées ou d'estimer les éventuelles données manquantes. L'outil centralise la collecte et l'accès aux rapports d'évaluation créés au cours du processus de vérification. Il est donc un support opérationnel garant de la conformité des données intégrées dans la base.

Au sein du portail *ESPON Database Portal*, l'outil est instancié trois fois en fonction du type de jeu de données, répondant ainsi à leur démarche de qualité respective. La figure 2 présente les étapes et les acteurs du processus de vérification des jeux de données de type *Key Indicators*.

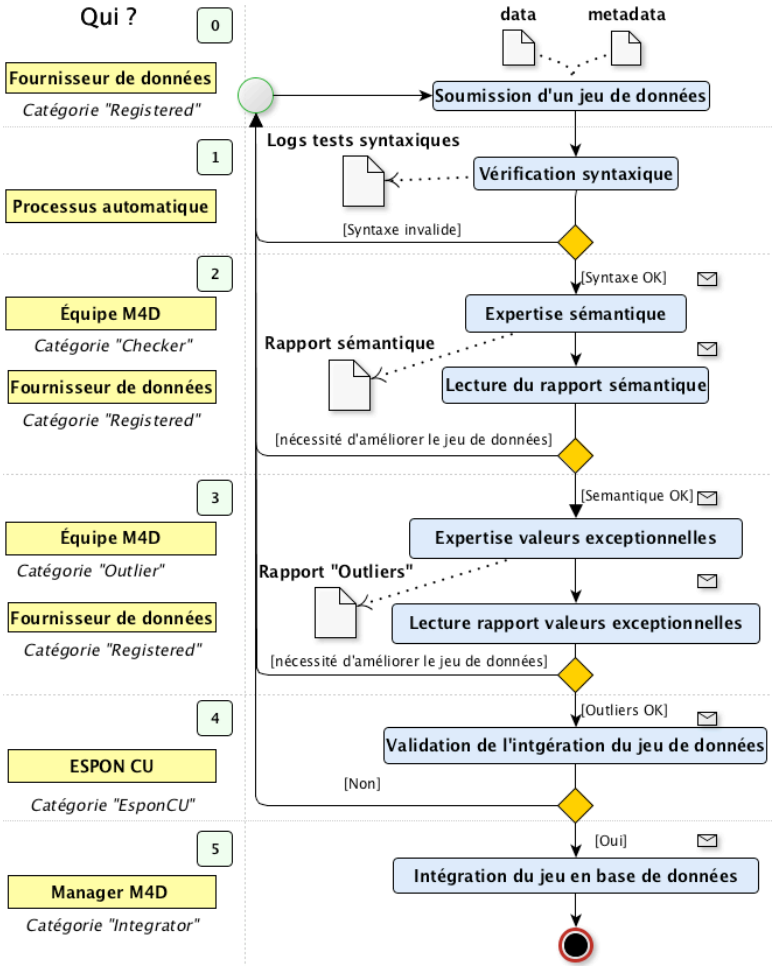

Figure 2 –Description des étapes et des acteurs du processus de vérification de jeux de données du type « Key Indicators »

Une vérification syntaxique et deux expertises sémantique et statistique permettent de contrôler un jeu de données du type *Key Indicators* (cf. étapes 1, 2, 3 figure 2). De ces étapes résultent des rapports d'évaluation et de recommandations à l'égard du producteur de données. Ces rapports forment les documents de référence pour l'évaluation de la qualité du jeu de données (Bergdahl et al., 2007). Le producteur de données peut consulter ces rapports au sein du *Checking tool* et visualiser l'état d'avancement des jeux de données grâce au tableau de bord présenté figure 3.

Dans l'application web *ESPON Database Portal* sont définies les catégories d'utilisateurs du *Checking tool* (*Registered, Checker, Outlier, EsponCU, Integrator, Administrator)* et les étapes auxquels ils doivent intervenir. Aux étapes sont associés des états : *non démarré, en cours, abandonné, effectué*. L'outil sollicite les acteurs par notification automatique, lors du passage d'un état à un autre ou lorsque le jeu de données franchit une nouvelle étape du processus.

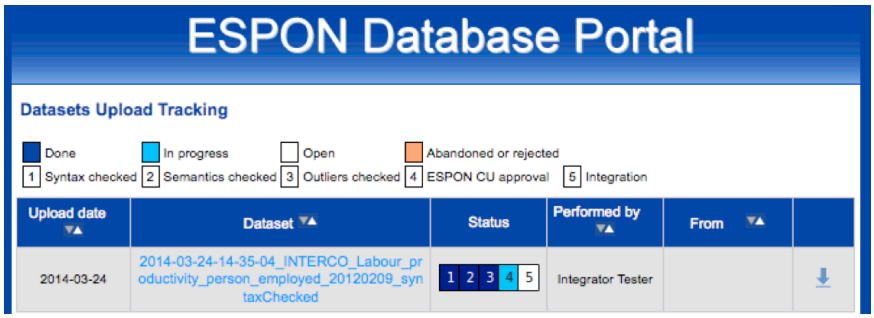

Figure 3 - Checking Tool - Interface de suivi des Key Indicators dataset.

La première étape du processus consiste à contrôler automatiquement la syntaxe du jeu de données. Le composant *Dataset Check* de *STedi* vérifie champ par champ le fichier soumis par le TPG afin de déterminer, entre autres, si tous les champs obligatoires sont présents et remplis. Un rapport automatique est généré, il comporte l'ensemble des traces d'erreurs détectées. Les messages d'erreur aident le fournisseur de données à conformer son jeu au modèle attendu. Le composant *Dataset Check* est intégré à l'outil *Checking tool* en ligne. Il est également disponible au téléchargement depuis le portail, en exécutable autonome, afin que les TPG puissent tester leurs jeux de données hors ligne.

Le TPG corrige le jeu de données jusqu'à ce qu'il ne comporte plus d'erreurs syntaxiques (figure 2, étape 1). Alors, l'étape de vérification sémantique est enclenchée par l'envoi d'une notification automatique à l'expert en charge de cette analyse. L'expert consigne dans un rapport ses observations concernant la qualité sémantique des métadonnées. Il vérifie notamment que les métadonnées relatives à la méthodologie employée pour la construction du jeu sont compréhensibles par tous. Le rapport sémantique est déposé par l'expert et accessible aux acteurs concernés depuis l'interface du *Checking tool*. Il contient par exemple une recommandation adressée au producteur de données (« *Ensure the reproducibility of such typologies [...,] add to the datasets delivered the input indicators used to calculate the typologies and [...] precise the methodology) »)*. L'expert pointe les champs de métadonnées devant être davantage détaillés ou corrigés et propose des solutions au producteur de données pour améliorer la qualité du jeu. Suite à la lecture du rapport, le TPG procède aux modifications à apporter (figure 2, étape 2).

A l'étape 3 intervient un expert statisticien en charge de détecter la présence de valeurs exceptionnelles (*outliers*) dans le jeu de données. Une batterie de tests semi-

automatiques (méthodes statistiques d'analyse spatiale, aspatiale[9], distribution des données, valeurs manquantes) est exécutée sur le jeu de données, en fonction de la nature des données de chaque indicateur (Charlton, 2012). Concernant les méthodes d'analyse spatiale, la mesure de l'Indice de Moran est par exemple utilisée pour quantifier la régularité spatiale d'un phénomène (autocorrélation spatiale) : la mesure négative d'un indice pour une zone indique que la valeur de l'indicateur pour cette zone est très différente des valeurs observées dans les zones immédiatement voisines, suggérant la présence d'une valeur exceptionnelle. L'expert produit un rapport de ses observations et le fournisseur de données vérifie si les valeurs exceptionnelles détectées sont attendues ou le fait d'erreur. Les acteurs s'impliquent et discutent de la qualité des données via l'outil.

A l'étape 4, le commanditaire du projet (ESPON Coordination Unit) valide ou non l'intégration du jeu de données dans la base de données. Sa décision s'appuie sur les deux rapports d'expertise soumis aux étapes précédentes.

L'outil, en ligne depuis Janvier 2013, a permis le suivi d'une centaine de jeux de données. Il coordonne et facilite l'interaction des différents acteurs pour améliorer la qualité des jeux de données au fur et à mesure des étapes. Disponibles au sein du portail web *ESPON Database Portal*, l'outil *Checking tool* est l'un des composants de l'infrastructure logicielle *STedi* présentée dans la section suivante.

5. « STedi » : des composants pour la gestion de l'ensemble du flux de données

L'infrastructure logicielle *STedi* repose sur une architecture « composant » qui intègre les outils présentés précédemment (*Checking Tool, Dataset Check,* etc.). La base de données relationnelle est elle-même un composant de l'infrastructure. Elle implémente le modèle de données présenté dans cet article grâce au SGBDR PostgreSQL-PostGIS. L'infrastructure logicielle établit les liens entre différents composants pour la gestion du flux d'information dans son ensemble. La figure 4 présente cette infrastructure logicielle et l'ordre d'intervention des composants pendant la gestion du flux de données.

Le composant SUNI (pour *Statistical Units Nomenclature Integrator*) permet d'intégrer des nomenclatures territoriales dans une base de données relationnelle reposant sur le modèle de données présenté (figure 4, encadré 1). Le composant *Checking tool* (figure 4, encadré 2) inclut d'autres composants tels que le composant *Dataset Check,* chargé de vérifier la conformité syntaxique d'un jeu de données. Le composant *Dataset Integrator* permet d'intégrer un jeu de données dans la base de données (figure 4, encadré 3), il convertit des éléments du jeu de données, délivré au format XLS, en objets Java manipulables. Le composant *Data Access* offre des

[9] Les méthodes statistiques employées pour la détection d'*outliers* sont spatiales ou aspatiales selon la pertinence de considérer ou non la disposition spatiale des unités statistiques de l'indicateur.

fonctionnalités de recherche pour accéder aux données dans la base. Ce composant peut être utilisé par un autre composant du type *web application* pour accéder de manière spécifique aux données (figure 4, encadré 4). Il intervient aussi lors de la restitution de données via des web services OGC (*Open Geospatial Consortium*) du type WFS (*Web Feature Service*) ou WMS (*Web Map Service*) (*cf.* encadré 5, figure 4).

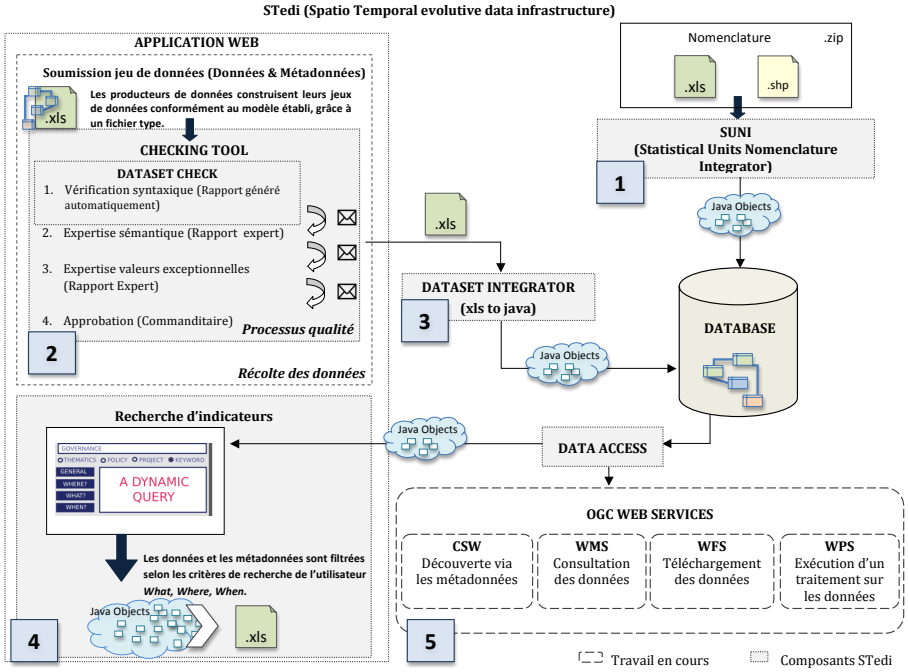

Figure 4 - Vue générale de l'infrastructure logicielle STedi

En termes d'ingénierie logicielle, le système *STedi*, écrit en langage Java, repose sur la notion de module telle qu'elle est entendue par le modèle de projet Maven. Les modules Java (par exemple *st-metadata, st-spatial, st-utils*) permettent de créer les composants de l'infrastructure *STedi* tels que le *Dataset Check* ou le *Dataset Integrator*, outils mobilisables indépendamment les uns des autres.

L'infrastructure implémentée dans le cadre du programme ESPON est une instance de l'infrastructure logicielle *STedi*. Au sein de cette instance, le composant *web application* répond aux besoins spécifiques des commanditaires du programme ESPON, notamment en ce qui concerne l'interface de recherche. Les autres composants dépendent moins des spécificités du programme ESPON, ils apportent des solutions pour la gestion de données. Les outils peuvent ainsi répondre aux problématiques d'organismes ayant les mêmes préoccupations mais dans un autre

domaine : conformer les fournisseurs à un modèle de données, veiller à la qualité des données, les intégrer dans une base de données relationnelle, etc. L'élaboration modulaire de l'infrastructure est un gage de son potentiel d'évolutivité, car l'architecture peut être modifiée pour rester utilisable malgré des pratiques ou des supports (nomenclatures, données) qui évoluent. Des composants nouveaux peuvent facilement intégrer l'infrastructure. Le système *STedi* a été construit en veillant à l'évolutivité des composants, ainsi la base de données supporte l'évolution des nomenclatures du territoire et l'outil *Checking tool* peut être adapté à des démarches de qualité qui évoluent (par exemple, des étapes supplémentaires de vérification peuvent être définies).

L'infrastructure pourrait, par son caractère modulaire, être implémentée dans le cadre de projet citoyen collaboratif, par exemple. Des utilisateurs proposeraient, au sein de la plateforme, des données historiques concernant leur commune. Le composant SUNI intégrerait différentes versions de nomenclatures définies au niveau de la commune, à différentes époques, donc selon différents découpages territoriaux. Le composant *Dataset Check* veillerait à l'harmonisation des données. Des experts s'assureraient, par le biais du composant *Checking tool,* de la qualité des données recueillies, en vérifiant les sources notamment. Le composant *Checking tool* aiderait aussi à modérer les dépôts. Les utilisateurs, par le biais des expertises, pourraient juger de la qualité des données.

6. Conclusion et perspectives

Le modèle de données créé pour les données statistiques territoriales est conforme aux normes et la dimension spatiale du modèle permet de représenter les évolutions des unités statistiques au cours du temps, comme le recommande l'annexe 3 de la directive INSPIRE (Commission Européenne, 2013). Le système opérationnel *STedi* assimile le modèle proposé et permet la diffusion de métadonnées riches par le biais des standards préconisés. Si le recours à ces standards ainsi que l'architecture modulaire confèrent à l'infrastructure un caractère générique et transposable à d'autres domaines, la combinaison d'un modèle étendu aux données statistiques territoriales et d'outils aidant à la vérification de la qualité des données font de *STedi* une infrastructure singulière. Le modèle de données pourrait encore être étendu pour inclure des métadonnées relatives à la qualité des données telles que celles proposées par le modèle EURO-SDMX. Concernant la démarche de qualité des données, les rapports d'évaluation créés au cours du suivi mis en place par le *Checking tool* pourraient être diffusés, non seulement aux parties prenantes, mais aussi aux utilisateurs, afin qu'ils identifient, par exemple, les valeurs exceptionnelles d'un indicateur. Quant à l'accès aux données, nos travaux actuels consistent en l'implémentation de services web respectant les spécifications de l'OGC. Ils permettront de garantir l'interopérabilité de l'infrastructure *STedi*, telle que préconisée par la directive INSPIRE et lui confèreront la qualité d'Infrastructure de Données Spatiales. Enfin, la description du modèle de données et métadonnées selon

un modèle de graphe RDF immergé dans les technologies du web sémantique permettra d'indexer et d'améliorer l'accessibilité des données statistiques territoriales par le web des données liées (ou *Linked Open Data Cloud*).

Bibliographie

Bergdahl M., Ehling M., Elvers E., Földesi E., Körner T., Kron A., Lohauß P., Mag K., Morais V., Nimmergut A., Viggo Sæbø H., Timm U., João Zilhão M., (2007). *Handbook on Data Quality Assessment Methods and Tools*, Rapport technique Eurostat, 2007.

Bretagnolle A., Guérois M., Averlant G., Mathian H., Delisle F., Lizzi L., Giraud T., (2011). *Naming UMZ, a database now operational for urban studies.* Rapport technique ESPON 2013 Program, mars 2011.

Charlton M., Harris P., Caimo A., (2012). *Detecting and handling anomalous data in M4D*. Rapport technique ESPON 2013 Program, juin 2012.

Commission Européenne (2013). *Annexe IV*. Journal Officiel de l'Union Européenne L331, décembre 2013.

European Commission Joint Research Centre, (2013). *INSPIRE Metadata Implementing Rules : Technical Guidelines based on EN ISO 19115 and EN ISO 19119*. Rapport technique, novembre 2013.

Götzfried A., (2009). *SDMX The basis for renovating the ESS Metadata Systems.* Présentation Eurostat, 2009.

IGN (2012). *Présentation INSPIRE*, http://inspire.ign.fr/directive/presentation.

OCDE, (2002). *Principaux indicateurs économiques – Analyse méthodologique comparative : indices des prix à la consommation et des prix à la production. Supplément 2*. Publication OCDE, 2002.

Plumejeaud C., Gensel J., Villanova-Oliver M., (2010). Opérationnalisation d'un profil ISO 19115 pour des métadonnées socioéconomiques. *Acte du colloque INFORSID 2010*, Marseille.

Plumejeaud C., (2011). *Modèles et méthodes pour l'information spatio-temporelle évolutive*. Thèse en Informatique, Université de Grenoble.

Sadou-Harireche N., (2007). *Evolution Structurelle dans les Architecture Logicielles à base de Composants*. Thèse en Informatique, Université de Nantes.

Telechev A., Le Rubrus B., (2013). *ESPON Data and Metadata Specifications*. Rapport technique ESPON 2013 Program, février 2014.

Ysebaert R., Salmon I., Le Rubrus B., Bernard C., (2014). Recueil, traçabilité et restitution des données territoriales du programme ESPON. *Acte du colloque Fronts et frontières des sciences du territoire 2014*, Paris.

Ysebaert R., Le Rubrus B., (2014). *How to deliver my data?* Rapport technique ESPON 2013 Program, février 2014.

Ce travail a été réalisé dans le cadre du projet 2013/3/3 ESPON Database and Data Development – Phase II du Programme Opérationnel ESPON 2013