

UTILISER LES MAJUSCULES D'IMPRIMERIE

A REMPLIR PAR L'ETUDIANT

UNIVERSITE ou
ETABLISSEMENT Telecom ParisTech
AUTEUR
NOM DE NAISSANCE ASSAF
PRENOM(S) Ahmad
NOM D'USAGE Ahmad Assaf
NOM FIGURANT SUR LA PUBLICATION, si différent des deux précédents
DATE DE NAISSANCE (*) (Format JJMMAAAA) 09081986

ADRESSE(*)
N° et RUE 6 Boulevard Dugommier
CODE POSTAL 06600 Pays France
VILLE Antibes
ADRESSE ELECTRONIQUE(*) (facultatif) ahmad.assaf@eurecom.fr

DATE DE SOUTENANCE ANNEE (AAAA) 2015 MOIS (MM) 12

DIRECTEUR DE THESE
SPECIALITE Multimedia & Semantic Web
NOM Troncy
PRENOM (S) Raphaël

TYPE DE DOCTORAT : COCHER LE TYPE DE DOCTORAT

☒ THESE DOCTORAT ☐ THESE D'EXERCICE ☐ THESE DOCTORAT D'ETAT

DISCIPLINE Semantic Web and Information Retrieval

INDICATIONS BIBLIOGRAPHIQUES :

Nombre de volumes : 1 Nombre total de pages : 219 Nombres de références bibliographiques : 151

AUCUN TEXTE MANUSCRIT N'EST ADMIS. DACTYLOGRAPHIER OU COLLER LE TEXTE IMPRIME DANS LES ZONES PREVUES.

TITRE DE LA THESE EN FRANÇAIS : Transcrire en toutes lettres les symboles spéciaux.

Fourniture automatisée de données par l'enrichissement sémantique pour l'informatique décisionnelle

(*) Les informations demandées sur ce formulaire sont destinées, à l'exception de l'adresse, à la constitution du catalogue collectif du Sudoc. Le Sudoc a fait l'objet d'une déclaration à la CNIL (arrêté du 10 septembre 1991 : publié au Journal Officiel du 1^{er} octobre 1991). La déclaration est obligatoire en application de l'arrêté du 25 septembre 1985 relatif au dépôt, signalement et reproduction des thèses ou travaux présentés en soutenance en vue du doctorat. Le droit d'accès et de rectification prévu par la loi n° 78 – 17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés ne concerne que les informations de la première page de ce formulaire. Il peut s'exercer auprès de : Agence bibliographique de l'enseignement supérieur – BP 4367 / 34196 MONTPELLIER CEDEX 5.

RESUME DE LA THESE EN FRANCAIS

Chaque cadre doit contenir un résumé de 1700 caractères maximum, espaces compris. En cas de dépassement, la coupure sera automatique.

Le doctorant adressera son texte sous forme électronique (disquette, etc.)

Pour les modalités pratiques, contactez votre bibliothèque.

La Business Intelligence (BI) classique et même les nouveaux outils de visualisation Agile concentrent une grande partie de leurs caractéristiques de vente sur des visualisations attrayantes et uniques. La préparation des données pour ces visualisations reste une tâche beaucoup plus difficile dans la plupart des projets BI, qu'ils soient grands ou petits. Le provisionnement des données en libre-service vise à résoudre ce problème en fournissant la découverte intuitive des ensembles de données, ainsi que l'acquisition et l'intégration des données techniques pour l'utilisateur final.

L'objectif de cette thèse est de fournir un cadre qui permet l'approvisionnement des données en libre-service dans l'entreprise. Ce cadre permet aux utilisateurs métiers de rechercher, inspecter et réutiliser les données par le biais de profils de jeux de données sémantiquement enrichies.

Dans cette thèse, nous proposons un mécanisme pour attacher automatiquement des méta-informations aux objets des données en tirant parti des bases de connaissances comme DBpedia et Freebase ce qui facilite la recherche et l'acquisition de données pour les utilisateurs professionnels.

En plus, nous proposons une approche évolutive automatique pour extraire, valider, corriger et générer des profils d'ensembles de données liées descriptives. Cette approche applique plusieurs techniques afin de vérifier la validité des métadonnées fournies et pour générer des informations descriptives et statistiques pour un ensemble de données particulier ou pour un portail de données entier.

RESUME DE LA THESE EN ANGLAIS

Classic Business Intelligence (BI) and even the newer Agile Visualization tools focus much of their selling features on attractive and unique visualizations. Preparing data for those visualizations still remains the far most challenging task in most BI projects large and small. Self-service data provisioning aims at tackling this problem by providing intuitive datasets discovery, data acquisition and integration techniques to the end user.

The goal of this thesis is to provide a framework that enables self-service data provisioning in the enterprise. This framework empowers business users to search, inspect and reuse data through semantically enriched datasets profiles.

In addition, we propose a scalable automatic approach for extracting, validating, correcting and generating descriptive linked dataset profiles. This approach applies several techniques in order to check the validity of the metadata provided and to generate descriptive and statistical information for a particular dataset or for an entire data portal.

We further present an extensible quality measurement tool implementing this framework that helps on one hand data owners to rate the quality of their datasets and get some hints on possible improvements, and on the other hand data consumers to choose their data sources from a ranked set.

TITRE DE LA THESE EN ANGLAIS : Transcrire en toutes lettre les symboles spéciaux

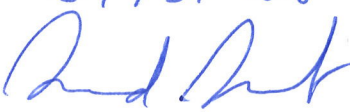
Enabling Self-service Data Provisioning Through Semantic Enrichment of Data

PROPOSITION DE MOTS-CLES : ces termes décrivant le contenu de la thèse seront saisis dans le Sudoc.

Semantic Web
Data Quality
Data Analysis
Data Enrichment

Data Profiling
Linked Data
Schema Matching
Open Data

DATE ET SIGNATURE DE L'AUTEUR

24 / 21 2016


A REMPLIR PAR LE SERVICE DE DOCTORAT

**UNIVERSITE ou
ETABLISSEMENT**

AUTORISATION DE REPRODUCTION PAR LE JURY

☐

OUI

☐

NON

☐

NON : thèse non reproduite car non corrigée dans les 3 mois réglementaires

☐

NON : thèse confidentielle JUSQU'EN

APRES VERIFICATION DATE et CACHET DU SERVICE DE DOCTORAT

A REMPLIR PAR LA BIBLIOTHEQUE

CATALOGAGE DE LA THESE ASSURE PAR LA BIBLIOTHEQUE DE L'UNIVERSITE OU DE L'ETABLISSEMENT

LOCALISATION DE L'EXEMPLAIRE ORIGINAL

RBCCN NUMERO D'IDENTIFICATION

Composé de 12 caractères : les 4 premiers indiquent l'année de soutenance, les 4 suivants définissent l'établissement, les 4 derniers correspondent à un numéro séquentiel

DISCIPLINE (telle qu'elle est indiquée sur la page de titre de la thèse.)

INDEXATION RAMEAU

1

2

3

4

5

INDEXATION LIBRE

RAPPEL : SI LE JURY A AUTORISE LA REPRODUCTION DE LA THESE, UN EXEMPLAIRE DE LA THESE ET DU FORMULAIRE SERONT TRANSMIS A L'ATELIER NATIONAL DE REPRODUCTION DES THESES (ANRT) DE GRENOBLE OU DE LILLE.

DATE ET CACHET DE L'ETABLISSEMENT