Thomas FRANCART
Tours, Paris
(+33) 06 71 11 25 97

38 ans, 1 enfant

thomas.francart@sparna.fr

Expert Web de Données

Conseils – Développement JEE – Formations
Ontologies, Thesaurus, Organisation des Connaissances
15 ans d'expérience

Parcours

La satisfaction clients, l'augmentation des demandes et l'embauche d'un salarié sont les signes de la réussite de mon activité.

Consultant Indépendant – Scparna *(depuis aout 2012)*

Conseil, développement, formations autour du web de données, des systèmes d'organisation de connaissances, et des architectures d'annotation/indexation de contenus. Nombreux clients d'envergure nationale et internationale (Europe, UNESCO, Etat du Luxembourg, INSEE, ISSN, Huma-Num, etc.)

Embauche d'un salarié en septembre 2016.

Conseil :

- Expert technique dans le projet européen ELI (European Legislation Identifier) : formations des états membres, modélisation d'ontologie, standardisation schema.org, analyses, alignement de vocabulaires ;
- Etat du Luxembourg : spécifications et architecture du nouveau portail législatif Legilux, publication des vocabulaires en SKOS ;
- INSEE : assistance à la finalisation du standard XKOS (extension de SKOS pour les classifications statistiques) ;
- ISSN : assistance à la modélisation du graphe de données ISSN ;
- Anaphore : modélisation d'une ontologie de description des ressources archivistiques ;
- Réseau Canopé : modélisation et diffusion des vocabulaires ScolomFr ;
- INRA : reprise de données et intégration dans l'outil VIVO ;

Développement et intégration de solutions :

- UNESCO : portail sémantique et intégration de VocBench et Skosmos pour la publication du thesaurus de l'UNESCO ;
- IRSTEA : intégration de VocBench et Skosmos pour la publication du thesaurus de l'IRSTEA ;
- ELI : validateur de métadonnées ELI basé sur SHACL ;
- Office des Publications de l'Union Européenne : développement de composants d'éditions et de validation de données RDF ;
- Huma-Num : développement de la plate-forme Nakala d'exposition de données en SHS ;
- Service SKOS Play de visualisation, test et conversion de thesaurus;

Formations web de données : Formateur web sémantique pour l'ADBS depuis 2011 (2 sessions par an) ; Formations à l'INHA, à l'INIST, au CNRS (réseau FRANTIQ), à Canopé Poitiers, à l'ABES, à l'IRSTEA, au SANDRE Limoges, Editions Francis Lefevre, etc.

Notamment grâce à mon implication, l'effectif de Mondeca est passé de 6 personnes en 2003 à 22 en 2011.

Directeur Technique / CTO – Mondeca *(octobre 2007 – juillet 2012)*

Responsable du produit ITM, logiciel de gestion/modélisation de connaissances et de vocabulaires, orienté web sémantique, de l'éditeur Mondeca (Paris, 22 personnes).

Définition des orientations du produit, déblocage des verrous technologiques, analyse des besoins client et intervention comme expert dans des projets internationaux et des projets R&D, forte implication dans l'avant-vente. Mise en place des processus qualité.

Design des architectures d'intégration entre les outils de gestion des systèmes d'organisation de connaissances (Mondeca), d'annotation de contenus (Temis, Arisem, Gate), de moteurs de recherche (SolR, Antidot, Exalead), et de bases RDF (Sesame).

Expertise reconnue dans les technologies du web sémantique (RDF, OWL, LOD…).

En mars 2004 j'ai repris seul la responsabilité du logiciel ITM

Architecte logiciel J2EE – Mondeca *(avril 2004 – septembre 2007)*

Responsable d'une équipe de 4 personnes pour le développement de la solution ITM. Analyses fonctionnelles, techniques, architectures J2EE, suivi de l'équipe, intégration de composants externes (text-mining, moteurs de recherche).

Responsable du suivi, de la réalisation et de l'aboutissement des premiers projets internationaux de Mondeca (USA, Royaume-Uni, Belgique). Nombreuses interventions sur des projets d'ingénierie documentaire pour l'édition, l'industrie, le tourisme ou la défense.

Développeur J2EE – Mondeca *(février 2003 – mars 2004)*

Développement des interfaces web du logiciel ITM : navigation – édition – recherche. Technologies J2EE.

Stagiaire développeur – UTT Troyes *(Septembre 2001 – Février 2002)*

Développement d'une "place de marché électronique", basée sur les standards de représentation de connaissances Topic Maps. Technologies Java, XML.

Formation

University of Pennsylvania – Philadelphia – USA *(2000-2001)*

Echange universitaire d'un an avec l'University of Pennsylvania, Philadelphia (Engineering school et Business school).

UTC – Université de Technologie de Compiègne *(1998-2000 & 2001-2003)*

Diplôme d'Ingénieur en Génie Informatique obtenu en octobre 2003.

Baccalauréat Scientifique *(1998) :* Mention Très Bien

Savoirs

Savoir-être

Autonome et indépendant, coopératif et réactif, ouvert et volontaire, inventif et réfléchi, animateur d'équipe et pédagogue, multi-tâches et organisé.

Savoir-faire : technologies et outils

Web de données : RDF, OWL, SPARQL, SKOS. RDF4J, SHACL, GraphDB, Virtuoso, Jena, RDF4J, SolR, Oracle Semantic Technologies, et tout l'écosystème d'outils open source.

J2EE – de la base de données à l'interface web : SQL, JDBC, EJBs (entités, session, messages), Hibernate, design patterns, JMS, web services, Spring, servlets, JSP, Struts, XSLT, GWT, HTML, CSS (bootstrap).

Serveurs d'applications : JBoss, Weblogic, Websphere, Tomcat, Jetty

Base de données : Oracle 11g, 10g, PostgreSQL, MySQL.

Développement : Git & SVN, Ant & Maven, Eclipse, Hudson, Artifactory, Sonar, JIRA.

Savoir-faire : organisation des connaissances

Manipulation de nombreux systèmes d'organisation de connaissances : ontologies, thesaurus, schémas de métadonnées, alignements, bases de connaissances, taxonomies, listes contrôlées, glossaires. Notamment : thesaurus de l'OMT [tourisme], de l'UNESCO [education], de l'IRSTEA [environnement], Eurovoc [Europe], GEMET [développement durable], taxonomie Wand [produits], catégories IPTC [presse], Code Officiel Géographique de l'INSEE et Geonames [géographie], DISCO [emploi], SNOMED et CIM-10 [médical] ...

Travail en particulier sur les problématiques suivantes : navigation, import, export, synchronisation, suivi des modifications, droits d'accès, dépréciation, modifications en masse, reporting, etc.

Travail important sur la gestion et la publication du thesaurus Eurovoc (eurovoc.europa.org) pour le compte de l'Office des Publication de la Commission Européenne.

Expertise sur les modèles de données importants FRBR, FRBRoo, CIDOC-CRM, schema.org, SKOS et sur la galaxie des ontologies du web de données. Traducteur français de la norme SKOS.

Langues : Anglais : courant, Allemand : pratique, Espagnol : notions

Divers

Plusieurs interventions dans les cursus universitaires (Master esDOC Poitiers, UTC, Paris X, Telecom SudParis & Brest, Université Pierre et Marie Curie).

Animation d'un blog **http://blog.sparna.fr** sur le web de données. Ancien contributeur de "leçons de choses", blog sur les activités de Mondeca.

Pratique de la philosophie : suivi de cours de philosophie et animation de soirées débat dans l'association Sesame à Paris XIXème.

Nombreux voyages à l'étranger, pratique de la photographie, de la course à pied (semi marathon) et de la randonnée en montagne.

Annexe : Quelques réalisations

ELI : European Legislation Identifier

Modélisation, Analyses, Développement, Formation, Standardisation

2014 - présent

L'initiative Européenne ELI (European Legislation Identifier), vise à utiliser les technologies du web de données pour la diffusion et la mise en lien des métadonnées des lois des Etats-Membres de l'UE. L'objectif est un accès et un échange facilité des données concernant les lois des Etats-Membres.

J'interviens depuis 2014 auprès du groupe de travail ELI comme expert technique sur divers aspects.

- Modélisation de l'ontologie ELI, basée sur FRBRoo ;
- Formation aux Etats-Membres sur le web de données et ELI (Irlande, Italie, France, Luxembourg, Danemark, Finlande, Autriche) ;
- Développement d'outil d'assistance à la publication des métadonnées ELI : ELI-validator ;
- Définition et proposition d'une extension « legislation » au vocabulaire schema.org ;
- Rédaction d'un guide de bonnes pratiques techniques pour l'implémentation de ELI, français/anglais ;
- Analyses et spécifications techniques sur les formats XML, la publication ELI en Open-Data, l'utilisation de JSON-LD ;

ISSN International Center

Modélisation et documentation

2017

Le Centre International de l'ISSN est un organisme international qui assigne et qui gère l'identification et la description des ressources périodiques (revues). L'ISSN souhaitait monter en compétences sur le web de données et les technologies associées, et modéliser le graphe de ses données publiques pour son nouveau portail https://portal.issn.org/.

La mission a consisté, à la suite d'une formation initiale de 2 jours, à formaliser et documenter le profil d'application des données de l'ISSN, en intégrant des propriétés des modèles schema.org, Dublin Core, BIBO, MARC21, FOAF, et d'autres.

Luxembourg

Spécification de la plate-forme de publication des données législatives du Luxembourg

2016-2017

Le Service Central de Législation du Luxembourg assure l'édition du Journal officiel du Grand-Duché de Luxembourg et la consolidation de la législation. Il a souhaité, dans la continuité de l'initiative ELI, se doter d'un nouveau portail de diffusion de la loi : legilux.public.lu. Ce portail est entièrement basé sur des données sémantiques en RDF, diffusées dans la plateforme Casemates.

La Phase 1 du projet a consisté dans la spécification de la nouvelle architecture de diffusion des données législatives au travers de Casemates. Le système comprend un repository documentaire (HTML ; PDF et XML des textes de loi) associé à une base de métadonnées RDF, sur un modèle dérivé de FRBR et proche de celui de ELI.

A la suite du développement de Casemates, j'ai accompagné dans la Phase 2 le SCL dans la spécification des fonctionnalités attendues dans le nouveau portail Legilux. L'architecture finalement retenue est un couplage RDF+ElasticSearch.

La Phase 3 a consisté à mettre en place la plateforme Skosmos pour la diffusion de tous les vocabulaires controlés utilisés dans la description des lois.

Réseau Canopé

Publication et versionnement de thesaurus en SKOS

2016-2018

Le Réseau Canopé édite des ressources pédagogiques transmédias (imprimé, web, mobile, TV), répondant aux besoins de la communauté éducative. A ce titre il maintient et publie les vocabulaires permettant l'indexation de ces ressources : le ScolomFr.

La Phase 1 du projet a consisté à définir un profil d'application basé sur SKOS pour la diffusion du ScolomFr, en prenant en particulier en compte l'évolution du ScolomFr, avec gestion des versions et des dépréciations de Concepts.

La Phase 2 a consisté à intégrer la plate-forme de diffusion de thesaurus Skosmos pour diffuser le ScolomFr. Ici encore, l'aspect temporel était prépondérant, en permettant d'afficher et de naviguer dans les versions successives du vocabulaire.

UNESCO

Gestion et publication de thesaurus en SKOS

2015-2016

L'UNESCO dispose d'un thesaurus en 4 langues (français, anglais, espagnol, russe), de 4400 concepts. L'organisation souhaitait remettre à plat son système de gestion de thesaurus en le rendant compatible avec les standards du web de données. L'objectif est de valoriser le thesaurus, de permettre sa réutilisation, et de le traduire dans un outil coopératif avec un système de workflow de validation.

Nous avons proposé une réponse au cahier des charges qui utilisait uniquement des composants open-source (VocBench, Skosmos, Jena, Skos Play), permettant ainsi de réaliser le système à moindre coût et à l'état de l'art par rapport aux bonnes pratiques de publication des données sur le web.

On pourra trouver plus de détails dans l'article http://blog.sparna.fr/2017/02/06/unesco-thesaurus-published-with-semantic-web-standards-and-open-source-software/

OPOCE

Office des Publications de la Commission Européenne

Edition – Luxembourg

2013-2014

L'Office des Publications de l'Union Européenne (basée au Luxembourg) assure l'édition des publications des institutions des Communautés européennes et de l'Union européenne (UE).

Mission de développement (60 jours) onsite et offsite, pour développer une interface de visualisation et de modification de données dans le Cellar, la base de métadonnées centrale au cœur du processus de production de l'Office.

Utilisation de Jena et RDForms, traitement de données dans le modèle FRBR. Contexte international et multiculturel.

Mission de spécifications (50 jours) pour l'implémentation de ELI (European Legislation Identifier) dans le portail législatif Européen http://eur-lex.europa.eu : analyses d'impact dans les différents systèmes du SI, rédaction de spécifications détaillées, réalisation de prototypes.

CNRS Huma-Num : Nakala

Recherche & Open-Data

2013-2014

La Très Grande Infrastructure de Recherche Huma-Num vise à faciliter le tournant numérique de la recherche en sciences humaines et sociales. A cette fin cette émanation du CNRS met un œuvre un dispositif humain (concertation collective) et technologique (services numériques pérennes) à l'échelle nationale et européenne en s'appuyant sur un important réseau de partenaires et d'opérateurs.

Huma-Num a confié à Sparna la réalisation de la plateforme Nakala (mise en ligne mi-2014) qui offre des services d'accès aux données elles-mêmes et des services de présentation des métadonnées. Les producteurs de données numériques, soulagés de la gestion purement technique, peuvent ainsi se consacrer à la valorisation scientifique de leurs données.

Rédaction de la réponse à l'appel d'offre, architecture technique, choix des composants, suivi de projet (relation client), suivi d'un développement outsourcé (Serbie), développement d'un connecteur OAI-PMH relié à une base SPARQL (Virtuoso).

JobTransport

2013

1ère mission significative en indépendant : de A à Z, l'enrichissement sémantique d'un moteur de recherche

Jobtransport.com est un site d'offres d'emplois spécialisé en transport et logistique. La mise en place du moteur de recherche SolR en remplacement d'une base relationnelle il y a 2 ans a apporté des bénéfices mais aussi une dégradation dans la précision de la recherche : il s'agissait donc d'améliorer SolR grâce à une ontologie.

Mise en place de tous les composants d'optimisation sémantique du moteur : création d'une ontologie, annotation automatique des contenus avec GATE, stockage des métadonnées dans une base RDF, alimentation de SolR avec l'ontologie et les métadonnées des contenus pour enrichir la recherche : extension sémantique sur synonymes, recherche à facette, amélioration de la précision, autocompletion.

Formation des équipes (niveau fonctionnel et niveau technique), design de l'architecture, paramétrage de GATE et Sesame, assistance à la création de l'ontologie.

Déblocage des verrous technologiques

2003 – 2011

Principal moteur d'innovation de Mondeca ; on peut citer notamment sur le plan technique :

- La conception et le développement de nombreuses fonctionnalités du logiciel ITM, parmi lesquelles l'affichage de graphes, le workflow de validation, les macros utilisateurs, le Single Sign On (serveur CAS) …
- La mise en place de tous les outils de développement logiciel pour l'équipe : de CVS à SVN, de Ant à Maven, Hudson, Artifactory, JIRA ;
- Des travaux d'optimisation des performances (jusqu'à x10) du moteur de requête d'ITM, et de portage sur de nombreux serveurs d'applications ;

Et également sur le plan des technologies sémantiques :

- La conception et le développement des modules de « dump » en RDF de la base de données d'ITM, de transformation de données RDF à base de règles, d'indexation de données RDF avec le moteur de recherche Lucene SolR, …
- La conception et le développement du produit CA-Manager de Mondeca, chaîne de traitement de documents basée sur UIMA, incluant un module de traduction de résultats du text-mining vers une ontologie ;
- La conception et le développement du produit Content Classifier de Mondeca, moteur de classification de contenus à base de règles SPARQL.

data.gouv.fr

Open Data – France

2011 – 2012

Etalab, mission du premier ministre chargée du développement de la plateforme française Open Data, a fait développer en 2011 le portail data.gouv.fr, point d'accès unique aux jeux de données ouverts par l'administration française.

Architecture technique de l'intégration du back-office, support aux équipes de Logica pour l'alimentation du moteur de recherche Exalead, expertise sur les aspects sémantiques, événements de communication (La Cantine).

bioMérieux

Santé – France

2011

Ce projet n'aura malheureusement pas abouti malgré mon implication dans une « gestion de crise »

Fabricant de matériel de laboratoire pour l'analyse et le diagnostic. Intégration d'ITM dans la solution Myla, système d'échange d'informations entre les outils de diagnostic du laboratoire, pour maintenir plusieurs terminologies médicales, incluant un mécanisme de gestion de versions. La gestion des correspondances entre versions d'une terminologie permet de faire communiquer entre eux des outils de diagnostic s'appuyant chacun sur une version différente.

Analyse des besoins client, architecture globale de la solution, développement d'un composant de calcul de différences entre versions de terminologies.

Mediapages

Médias – Québec

2011

Création du site de pages jaunes québécoises bilingues http://trouvetout.ca. Utilisation des solutions de Mondeca pour construire et maintenir la base des mots-clés associés à chaque professionnel dans l'annuaire, et suggérer de nouveaux mots-clés depuis le front-office. Intégration avec les produits d'Apptus (moteur de recherche et création assistée de taxonomies), d'Arisem (analyse linguistique), et IT2Media (gestion de la facturation et des fiches de l'annuaire).

Présentation des solutions et des produits sur place à Montreal, définition de l'architecture technique, rédaction de spécifications sur les reprises de taxonomies géographiques et produits, support aux équipes de développement de Vidéotron.

Thalès et EADS

Défense et sécurité – France

2009 - 2011

Systèmes de veille et d'aide à la décision. Intégration dans les architectures logicielles de Thalès et d'EADS au travers des projets de recherche OSEMINTI et VIRTUOSO (http://www.virtuoso.eu).

Coordinateur des 2 projets de recherche pour Mondeca, intégration de l'outil ITM dans les architectures respectives de Thalès et EADS, formation aux utilisateurs et aux développeurs.

OPOCE

Office des Publications de la Commission Européenne

Edition – Luxembourg

Déploiement d'une solution de gestion de thesaurus permettant à l'OPOCE de maintenir le thesaurus Eurovoc (26 langues). La solution s'intègre dans un workflow de travail complexe (traducteurs, validateurs, comité de relecture) et permet une publication du thesaurus sur le portail http://eurovoc.europa.eu dans les standards du web sémantique (RDF, SKOS).

Phase 1

2008

Phase 2

Analyse des besoins, spécifications et coordination des développements, dont l'implémentation des fonctionnalités de workflow dans ITM ; implémentation et paramétrage des exports SKOS du thesaurus vers le portail.
Déploiement d'une application d'alignement de vocabulaires (ITM-Align). Cette application permet d'aligner semi-automatiquement des thesaurus, étendant ainsi les

2010

capacités des moteurs de recherche sur les contenus.

Spécifications fonctionnelles et techniques de l'application d'alignement, intégration de l'API d'alignement du laboratoire INRIA Exmo, réalisation d'une chaine d'indexation de données RDF à l'aide de Lucene SolR, intégrant la prise en compte des alignements.

Lexis Nexis

Edition – France et USA

Editeur juridique français. Refonte complète du système informatique éditorial, impliquant le déploiement d'ITM pour stocker d'importants volumes de métadonnées juridiques (y compris les renvois vers les textes de loi), l'analyse automatique des contenus à l'aide de text-mining, la mise en place d'un moteur de recherche sur tous les contenus avec recherche à facettes. La refonte de l'informatique éditoriale, à base d'ontologies, a permis non seulement de faciliter l'acquisition, la création, la diffusion et la réutilisation des contenus, mais aussi de constituer une base de connaissances juridique (thesaurus et renvois des textes), désormais un véritable actif de cet éditeur.

Phase 1

2006

Responsable du déploiement d'ITM, du développement d'une chaine d'acquisition de métadonnées des documents, intégration d'outils de text-mining (Temis), benchmarks et amélioration des performances.

Phase 2

2010

Seconde implémentation, et déploiement de la chaine d'acquisition de métadonnées de documents, et d'indexation plein-texte (Antidot), au data-center de Dayton, USA.

Intégration UIMA avec Temis, spécification et coordination de l'implémentation d'un composant de classification de contenu à base de règles (SPARQL) s'appuyant sur les métadonnées (RDF). Participation à toutes les phases du projet, depuis la rédaction de la réponse jusqu'au support au déploiement de l'équipe US.

Thomson Scientific

Edition – USA

2007

Editeur scientifique. Déploiement d'une solution d'annotation automatique de résumés d'articles scientifiques de 1930 à 1967 (environ 1,5 million d'articles), en partenariat avec Temis. L'annotation automatique de ce fond important a permis de le valoriser et de le rendre accessible au travers d'outils de recherche.

Coordination du projet pour Mondeca (réunions téléphoniques hebdomadaire avec l'équipe US), développement du composant d'annotation automatique à base d'ontologies, avec des capacités d'inférence terminologique, intégration des outils de Temis, formation des utilisateurs sur site aux USA (Philadelphie).

TAO

Transitiong Applications to Ontologies

Recherche – Européen

2006-2007

Projet de recherche européen ayant pour objectif la fourniture de méthodologies et d'outils logiciels pour faciliter le passage des applications monolithiques classiques vers des applications basées sur des ontologies. Ce projet a regroupé des partenaires universitaires et industriels Anglais (dont l'université de Sheffield, développeurs de la plate-forme de text-mining Gate), Français (Mondeca, Dassault Systèmes), Bulgares (Ontotext), Slovènes et Espagnols (ATOS).

Représentant de Mondeca dans le projet. Participation aux réunions et revues de projet, conception, développement et présentation du produit CA-Manager dans le cadre du projet, intégration de Gate.

Hachette Filipacchi Media

Edition – France

2004 - 2005

Editeur pour la presse magazine people. Déploiement d'une chaine d'acquisition de connaissances et d'annotation de contenu des flux presse (PQN, PHN).

Développement des connecteurs avec la base XML Xylème et des écrans de validation de l'acquisition de contenu.

Wolters Kluwer

Edition – Belgique

2004

Editeur juridique belge bilingue Néerlandais-Français. Déploiement d'une solution de TMS (Thesaurus Management System) incluant un module de génération d'index de publication, à partir des métadonnées d'annotation de chaque chapitre sur les thesaurus. La génération semi-automatique d'index facilite un travail auparavant entièrement manuel, et permet une réutilisation plus aisée des unités de contenu.

Coordination de l'implémentation et de l'intégration du TMS et de l'algorithme de génération d'index (3 développeurs), formation et support sur site (Belgique), portage sur Oracle Application Server, maintenance applicative.