# TYPEX: Typeful certified XML integrating language, logic, and data-oriented best practices

ANR-11-BS02-007

**PPS, Univ. Paris Diderot - LRI, Univ Paris Sud - Tyrex, INRIA Rhône Alpes**

26 février 2014

# Présentation de TYPEX

Le but : faire des avancées sur le typage des langages de programmation adaptés à la manipulation de données au format XML.

# Structuration

**Consortium:**

PPS (coordination). Expertise XML langages de programmation.
LRI. Expertise XML orienté données.
Tyrex (ex WAM). Expertise XML par logique basée sur les solveurs

**Direction:**

Coordinateur: Giuseppe Castagna (PPS)

Responsables de site: G. Castagna (PPS), V. Benzaken (LRI), N. Layaïda (Tyrex)

Responsables Tasks: G. Castagna (T1&T5, PPS), N. Layaïda (T2, Tyrex), M. Sozeau (T3, PPS), K. Nguyễn (T4, LRI).

Nombre de membres permanents: 11

Permanents actifs (participation $\geq 25\%$): **6** ($= 2$ par site).

# Collaboration synergétique face à un défi scientifique

## Défi: typage XML précis difficile avec une seule approche

Chaque partenaire possède une expertise internationalement reconnue dans un domaine différent :

1. logique et solveurs;
2. programmation fonctionnelle;
3. langages de manipulation de données.

## Objectifs:

- combiner les trois approches dans un cadre unique et formellement certifié (notamment en Coq)
- définir un cadre fondamental, des techniques et des outils, pour analyser statiquement, implémenter efficacement et optimiser les transformations XML.

## Méthodologie:

1. améliorer chaque approche par « cross-fertilization »
2. produire une approche unique/unifiée.

# Organisation

**Dans la proposition:** 5 tâches (Tasks) organisées en sous tâches (track); tous les sites participent à chaque Task.

- Task 1: Administration et coordination
- Task 2: Polymorphisme et solveurs.
- Task 3: Implémentation et certification des standards XML
- Task 4: Propriété de données et des transformations XML
- Task 5: Intégration

**Depuis:** une sixième tâche élargissant le champ des données considérées pour prendre en compte les évolutions technologiques et de recherche.

- Task 6: Nouveaux formats (JSON, RDF, ...) et nouveau paradigmes (NoSQL, langages dynamiques)

# Task 2: Polymorphisme et solveurs.

**Track 1.** Fonctions polymorphes pour la programmation XML.

- Définition de fonctions polymorphes en CDuce
- Définition de l'inférence locale de types

**Track 2.** Polymorphisme et types d'ordre supérieur pour solveurs

- Traitement de types polymorphes dans le solveur
- Traitement de transformation/types d'ordre supérieur dans le solveur

**Timeline** (4 pas, les deux premiers en parallèle):

1. Génération de contraintes de (sous-)typage
2. Synthèse de types polymorphes
3. Résolution systèmes de contraintes
4. Conception et implémentation de CDuce polymorphe

# Task 3: Implémentation et certification des standards XML

**Track 1.** Moteur XPath basé sur les automates d'arbres

- Définir un modèle d'automate d'arbres permettant l'évaluation d'XPath (avec axes arrières) en temps $O(|Document| \times |Automate|)$
- Fournir une implantation efficace de ce modèle

**Track 2.** Formalisation et certification d'outils XML

- Preuve de correction (au moins partielle) du logiciel de la **Track 1** en Coq
- Preuve en Coq de la complexité du logiciel **L1**

**Timeline** (3 pas, dépendence séquentielle):

1. Implantation des automates avec axes arrières
2. Preuve de l'implantation (1) en Coq
3. Preuve de propriétés du logiciel **L1** (utilisant les outils développés en (2))

# Task 4: Propriétés des données et transformations XML

**Track 1.** Moteur de requête efficace basé sur les types

- Langage avec motifs + navigation XPath
- Typage précis des axes arrières

**Track 2.** Annotation et vérification statique de XQuery

- Invariants et contraintes logiques pour XQuery
- Utilisation d'outils externes pour la vérification de programmes XQuery

**Timeline** (4 pas, 1-2 et 3-4 en parallèle):

1. Définition d'un calcul avec motifs et chemins, ainsi que du système de types associé
2. Implantation par ajout de la navigation arrière dans CDuce
3. Définition d'un langage d'assertions logiques pour XQuery
4. Vérification de programmes par de multiples outils

# Task 5: Intégration

Tâche ambitieuse et risquée: son succès dépend du succès total des trois autres tâches techniques

Résultats attendus:

Une proposition (critère minimal de réussite) ou un prototype (critère maximal) pour langage de manipulation de données de nouvelle génération, basé sur le filtrage par motif, le polymorphisme, les assertions logiques, un typage précis et une implantation efficace et certifiée.

# Task 6: Nouveau formats et paradigmes

**Prendre en compte l'émergence de nouvelles technologies:**

- Nouveaux formats de données:
  JSON, RDF
- Nouveaux langages de rêquetes:
  Langages NoSQL (JaQL, MongoDB, CouchDB) et pour RDF (SPARQL)
- Nouveaux paradigmes:
  Langages dynamiques (Javascript, R, Ruby, Python)

# Résultats obtenus par TYPEX

# Cadre général:

## Calendrier

**Durée: 44 mois** (début: janvier 2012, fin: septembre 2015)
**Aujourd'hui: M25** (56%)

# Cadre général:

## Calendrier

**Durée: 44 mois** (début: janvier 2012, fin: septembre 2015)
**Aujourd'hui: M25** (56%)

### Phase 1: Cross-fertilization.

5 publications communes : indicateur en deça de la cross-fertilization réelle car l'influence réciproque est très forte:

1. Le délivrable D.2.c par TYREX résoud de manière logique le problème du sous-typage de la théorie développée par PPS.
2. Le délivrable D.4.a par LRI et PPS, est basé sur une technique (l'utilisation de zippers pour modéliser les axes « arrières » de XPath) introduite et développée par TYREX pour leur solveur.
3. Le prototype D.3.b (Task 3 Track 1) est actuellement développé par 3 chercheurs (Nguyễn, Schmitt et Sozeau) qui appartiennent aux 3 partenaires du projet.

### Phase 2: Intégration

Début mars 2014

# Task 2:

## État d'avancement: 90%

1. La géneration de contraintes de (sous-)typage à été décrite dans [CNXILP14] (présenté à **POPL**).
2. Synthèse de types en cours de soumission: le rapport [JGL13] et a été implémenté dans le logiciel **L1** et une publication est en cours.
3. La résolution de systèmes de contraintes a été définie dans [GGL13,CNXILP14]. Elle permet l'inférence locale de types et aussi la reconstruction de types (à la ML).
4. CDuce polymorphe: [CNXILP14] définit une implémentation efficace des fonctions polymorphes (comparable à une execution polymorphe). Elle a été implantée dans le prototype **L6** et l'implémentation pour CDuce est en cours (logiciel **L7**).

# Task 3:

**État d'avancement: 50%**

1. Réimplantation complète du moteur de requête basé sur les automates d'arbres (logiciel **L2**). Support des axes arrières et ajout des jointures en cours
2. Début de formalisation de la sémantique d'XPath en Coq
3. Preuve « papier » de correction

   Il reste à « traduire » la preuve papier en Coq et étudier formellement certains aspects du solveur (en Coq).
4. Extensions accomplies de la logique et du solveur pour supporter les attributs et valeurs. Application: premières analyses statiques pour CSS [GLQ12] (présentées à **WWW**)
   $\rightarrow$ perspectives de concrétisation des résultats du projet sur des applications réelles et vers le marché (prototype préindustriel)

# Task 4:

## État d'avancement: 80%

1. **[CINB13]** présente un calcul et un système de types contenant XPath ainsi que les motifs CDuce
2. Le calcul est en cours d'implantation comme une extension de CDuce (logiciel **L7**).
3. Deux travaux (**[JGL12]** et **[OGL12]**) menés par TYREX étudient l'utilisation de solveurs (solveur de $\mu$-calcul ou solveur SMT Z3) pour la preuve de propriété logiques (indépendances des mises à jour et évolution de code XQuery)
4. De plus dans **[GLV12]** Tyrex développe une nouvelle approche par inférence arrière pour le typage de XQuery, et formalise le problème du typage des axes arrières dans **[GGL13]**

$\Rightarrow$ Il reste à réunir les deux sous-tâches en réutilisant les acquis du point 3 ci-dessus et en les intégrant à CDuce + chemin

# Task 5 (intégration):

En démarrage (démarrage prévu en M24, aujourd'hui: M25)

# Task 6:

## État d'avancement: nouvelle tâche non prévue initialement

- Adaptation des techniques du sous-typage sémantique aux langages NoSQL, notamment JaQL [BCNS13] (publié à **POPL**) et MongoDB [Hus14] (publié à **JFLA**) (**D6.a**).
- Inclusion de requêtes SPARQL pour RDF [EGL12] (publié à **AAAI** et **IJCAR**) et leur évaluation [CEGL12] (publié à **ISWC**, (**D6.b**)).
- Formalisation de JavaScript en Coq [BCFGMNSS14] (publié à **POPL**) et d'analyses statiques basées sur l'évaluation formelle [BJS14] (publiée à **JFLA**) (**D6.c**).
- Certification du modèle relationnel en Coq (pour servir de base à la certification des langages NoSQL) [BCD14] (publié à **ESOP**) (**D6.d**).

# Publications:

## Journaux internationaux (1):

[GL14] P. Genevès, N. Layaïda. Equipping IDEs with XML Path Reasoning Capabilities. In **TOIT'14**, ACM Transactions on Internet Technology, 2014.

## Conférences internationales (12):

[BCD14] V. Benzaken, E. Contejean and Stefania Dumbrava. A Coq Formalization of the Relational Data Model. In **ESOP'14**, European Symposium on Programming, 2014.

**[CNXILP14]** G. Castagna, K. Nguyễn, Z. Xu, H. Im, S. Lenglet, and L. Padovani: Polymorphic Functions with Set-Theoretic Types. Part 1: Syntax, Semantics, and Evaluation, In **POPL'14**, 41st ACM Symposium on Principles of Programming Languages, 2014.

**[BCFGMNSS14]** M. Bodin, A. Charguéraud, D. Filaretti, P. Gardner, S. Maffeis, D. Naudziuniene, A. Schmitt, G. Smith. A Trusted Mechanised JavaScript Specification. In **POPL'14**, 41st ACM Symposium on Principles of Programming Languages, 2014.

**[BCNS13]** V. Benzaken, G. Castagna, K. Nguyễn, and J. Siméon: Static and Dynamic Semantics of NoSQL Languages. In **POPL'13**, 40th ACM Symposium on Principles of Programming Languages, 2013.

[INP13 ] H. Im, K. Nakata, S. Park: Contractive Signatures with Recursive Types, Type Parameters, and Abstract Types, In **ICALP'13** 40th International Colloquium on Automata, Language and Programming. 2013

[CEGL13] M.W. Chekol, J. Euzenat, P. Genevès, N. Layaïda. "Evaluating and benchmarking SPARQL query containment solvers". In **ISWC 2013**: The 12th International Semantic Web Conference, 2013.

**publications multi-site** **publications mono-site**

[GL13] P. Genevès, N. Layaïda. XML Validation: Looking Backward – Strongly Typed and Flexible XML Processing are not Incompatible (demo). **WWW 2013** The World Wide Web Conference. 2013

[EGL12] M.W. Chekol, J. Euzenat, P. Genevès, N. Layaïda. SPARQL Query Containment under RDFS Entailment Regime. In **IJCAR 2012**, International Joint Conference on Automated Reasoning. 2012

[GLQ12] P. Genevès, N. Layaïda, V. Quint. On the Analysis of Cascading Style Sheets, In **WWW 2012**, The World Wide Web Conference, 2012

[CEGL12] M.W. Chekol, J. Euzenat, P. Genevès, N. Layaïda. SPARQL Query Containment. In **AAAI 2012**, 26th AAAI Conference on Artificial Intelligence, 2012.

[JGL12] M. Junedi, P. Geneves, N. Layaida, XML Query-Update Independence Analysis Revisited, **DocEng'12**, ACM Symposium on Document Engineering, 2012.

[OGL12] R. Oliveira, P. Geneves, N. Layaida, Toward automated schema-directed code revision, **DocEng'12**, ACM Symposium on Document Engineering, 2012.

## Conférences invitées (1):

[Cas12] G. Castagna. Type-checking union, intersection, and negation types, Conférence invité a **ITRS'12** 6th workshop on Intersection Types and Related Systems.

## Conférences nationales (1):

[BJS14] M. Bodin, T. Jensen, A. Schmitt. Pretty-big-step-semantics-based Certified Abstract Interpretation, **JFLA'14**. January 2014.

[Hus14] A. Husson. Une sémantique statique pour MongoDB, **JFLA'14**. January 2014.

## Articles soumis (3):

**[CINB13]** G. Castagna, H. Im, K. Nguyễn, and V. Benzaken: A Core Calculus for XQuery 3.0, 2014. (article soumis à une conférence internationale)

**[CNX13]** G. Castagna, K. Nguyễn, and Z. Xu: Polymorphic Functions with Set-Theoretic Types. Part 2: Local Type Inference and Type Reconstruction, July, 2014.(article soumis à une conférence internationale)

[GGL13] N Gesbert, P. Genevès, N. Layaïda: A Logical Approach To Deciding Semantic Subtyping – Supporting function, intersection negation and polymorphic types (soumis à une revue internationale)

## Rapports techniques (2):

[JGL13] Louis Jachiet, Pierre Genevès, Nabil Layaïda. Type Synthesis for the Logical Solver: an Approach based on Query Automata. http://typex.lri.fr/files/synthesis.pdf

[GLV12] Pierre Genevès, Nabil Layaïda, Christine Vanoirbeek. XQTC: A Static Type-Checker for XQuery Using Backward Type Inference. http://hal.inria.fr/hal-00757867

# Logiciels (7):

**L1:** XML Reasoning Solver Project:

- **a.** Synthèse de types (XPath), avancement: 90%
- **b.** analyse de code et inférence de type pour XQuery: 70%
- **c.** combinateurs (macros) pour la logique: 100%
- **d.** fonctions polymorphes et sous-typage logique: 100%

Diffusion: prototypes offline à diffusion controlée et démo online (web interface).

**L2:** Tree Automata Toolkit (TATOO): An XPath engine based on tree automata: Avancement 80% (implementation de Navigational XPath). Diffusion: Git Tatoo, LGPL.

**L3:** XML Concepts formalized in the Coq proof assistant: Avancement: 50%. Diffusion: Git, LGPL.

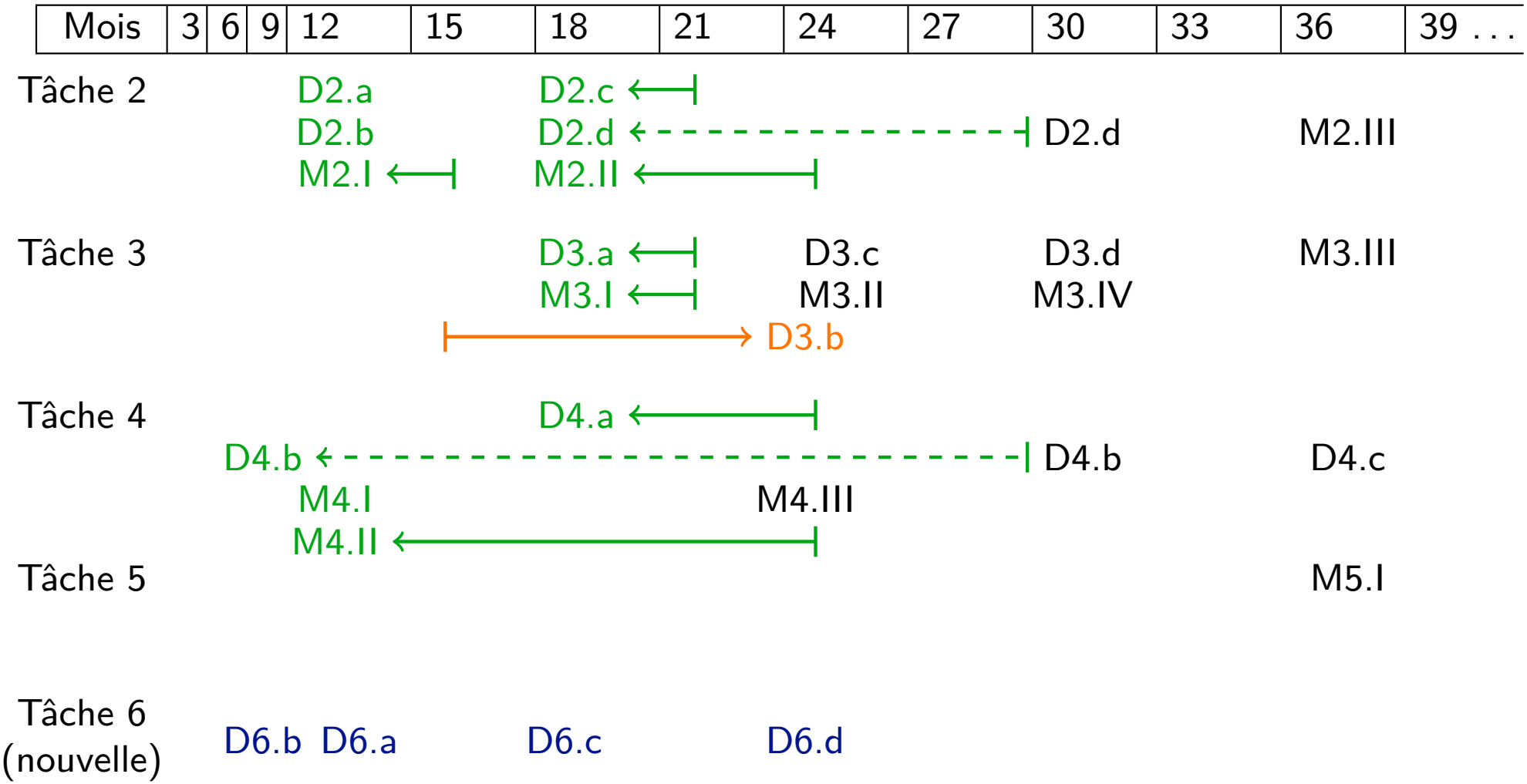
**L4:** Coq pour bases de données relationnelles: Avancement 100% (article ESOP 2014). Diffusion: bibliothèque Coq à diffusion controlée (on-demand).

**L5:** CDuce avec patterns navigationnels Avancement 80% (description). Diffusion: branche Git de CDuce.

**L6:** Fonction polymorphes avec types ensemblistes: Avancement 100% (prototype de test complet implémentant le papier POPL 2014). Diffusion: usage interne (on concentre tous les efforts de développement sur CDuce).

**L7:** CDuce polymorphe: Avancement 30% (sous-typage, inférence de types, achevé; compilation interfaçage avec OCaml démarré). Diffusion: branche Git de CDuce.

# Livrables et jalons



# Résultats marquants

1. Définition et développement d'une technique de sous-typage polymorphe à base de logique [GGL13]
2. Définition d'un calcul explicitement typé pour les types intersection [CNXILP13,Xu13]
3. Définition d'un modèle d'exécution efficace pour une extension polymorphe de CDuce [CNXILP13]
4. Inférence locale et reconstruction pour un système de types ensemblistes (avec union, intersection et négation) et récursifs : génération et résolution de systèmes de contraintes [CNX13]
5. Définition d'un modèle formel pour langages NoSQL [BCNS13]
6. Définition d'une sémantique formelle et d'un système de types pour XQuery 3.0 [CINB13] et en logique [GGL14]
7. Intégration de patterns "navigationnels" (à la XPath) dans CDuce [CINB13]
8. Définition et développement des premières techniques d'analyse statique de CSS [GGL12] et SPARQL [CEGL12a,CEGL12b,CEGL13]

# Faits marquants:

1. Pierre Genevès (Tyrex) vient de recevoir la **médaille de Bronze du CNRS** (voir `http://www.cnrs.fr/fr/recherche/prix/medaillesbronze.htm`).
2. L'article *Set-Theoretic Foundation of Parametric Polymorphism and Subtyping* par Giuseppe Castagna et Zhiwu Xu (PPS) a reçu la nomination par ACM SIGPLAN aux **CACM Research Highlights** (voir: `http://www.sigplan.org/Newsletters/CACM/Papers`). Il s'agit de l'article qui est à la base de la Task 2 de TYPEX et en particulier des livrables D.2.a, D.2.b, D.2.c et D.2.d.
3. Création de l'équipe Tyrex (CNRS, Inria, Université Grenoble Alpes, Grenoble Inp)

# Travail planifié

# A faire (nous sommes en M25)

Programme de recherche en ligne avec le projet soumis à l'ANR il y a 3 ans.

- Déliverable D2.d (M30): partiellement livré en avance (M18). Il reste à compléter l'extension de CDuce avec les fonctions polymorphes.
- Déliverable D3.d (M30): automated proof for the solver.
- Milestone M3.IV: the algorithm of the solver is verified.
- Délivrable D4.b (M30): partiellement livré en avance (M9). Il reste à intégrer les assertations.
- Milestone M2.III (M36): polymorphic CDuce.
- Milestone M3.III (M36): SXSI with backward axes is verified.
- Délivrable D4.c (M36): An XQuery engine equipped with a static type system and logical assertions.
- Délivrable M5.I (M36): fundamentals for the standardization of the XQuery next-generation language.

# Recrutements: Post-Docs

1. Sergueï Langlet (PPS) Fonctions Polymorphes pour la programmation XML [01/2012, 08/2012] Départ après 8 mois suite à son recrutement comme MdC à Nancy.
2. Hyeonseung Im (LRI) CDuce navigationnel et formalisation XQuery 3.0 [11/2012, 04/2014], puis Tyrex.
3. Pietro Abate (PPS). Implémentation du polymorphisme en CDuce [10/2013, 09/2014]

# Recrutements: Stagiaires

1. Adrien Husson (PPS et IBM TJ Watson New York) stage ENS Cachan: typage requêtes MongoDB [05/2013,08/2013]
2. Julien Lopez (PPS) stage EPITA: interface CDuce/OCaml pour types polymorphes. [02/2014, 08/2014]
3. Huibo Shi (LRI et Université Paris-Sud) : évaluation des performances de Tatoo (**L2**) face à l'état de l'art en matière de requêtes XML
4. Muhammad Junedi (Tyrex, en thèse depuis fin 2012): analyse d'indépendence de requêtes et mises à jour XQuery;
5. Louis Jachiet (Stage ENS Paris puis thèse tyrex dès sept. 2014): synthèse efficace de types pour XQuery;
6. Abdullah Abbas (Tyrex) : réécriture de requêtes SPARQL pour analyse statique
7. Martí Bosch (Tyrex) : analyse et optimisation automatique des CSS

# Réunions

[Programmes détaillés: `http://typex.lri.fr/meetings.html`]

**Plenaires:**

- [2012] Janvier 9-10. Lieu: Paris.
- [2013] Decembre 16-17. Lieu: Paris.
- [2014] Avril 14-15. Lieu: Grenoble.

**Thématiques:**

- 7/9/2012 Réunion de travail du Task 3 Track 2. Lieu: Paris. Sites participants: tous.
- 21/4/2013 Réunion du steering committee. Lieu: Grenoble.
- 22/5/2013 Réunion de travail du Task 3 Track 2. Lieu: Paris. Sites participants: tous.
- Les partenaires PPS et LRI se rencontrent regulièrement avec cadence hebdomadaire.

# Difficultés rencontrées.

- Problème: Problèmes de santé du coordinateur
  Action: replanification du travail et des réunions; prolongation de 8 mois du projet.
- Problème: départ anticipé Post-doc PPS (poste MdC)
  Action: Restructuration du budget restant pour offrir 12 mois de post-doc à un nouveau candidat (recruté en Octobre 2013).
- Problème: difficulté pour trouver des stagiaires.
  Action: élargissement des établissements où les stages sont proposés (actuellement: MPRI, Master Logique, ENS Lyon, ENS Cachan, EPITA ...?)
- Problème: abandon d'un doctorant PPS pour le privé (ingénieur Google).
  Action: abandon des aspects parallélisation. Réaffectation des aspects implémentation au deuxième post-doc recruté par PPS.

- Problème: mutation d'un des permanents (Alan Schmitt) à l'INRIA Bretagne.
  Action: Continuation de la collaboration et gestion des fonds de la part de l'INRIA Rhône Alpes.
- Problème: difficulté Tyrex à trouver des candidats post-doc dans le profil recherché.
  Action: Coordination avec LRI et PPS. Transfert de 4 mois de post-doc de Tyrex à LRI; éventuel (décision et details en mars 2014) recrutement et mutation dans Tyrex des post-docs LRI ou PPS à la fin de leur embauche ou nouveau recrutrement (14 mois).
- Problème: Track 1 de la tâche 3 demandant un important effort de développement (par K. Nguyen, impliqué dans d'autres tâches). Retard du début de la Track 2 de la même tâche
  Action: Replanification de la Track 2 (la Track 1 étant terminée)

# Budget et ressources

| | | Missions | Fonctionn. | Equipement | Personnel |
|---|---|---|---|---|---|
| PPS | Alloué | 27950 | 10074 | 3600 | 73476 |
| | Dépensé | 8936 | 194 | 1194 | 35583 |
| | Engagé | 2416 | 2616 | == | 36550 |
| | Pourcent | 41% | 28% | 33% | 98% |
| LRI | Alloué | 16000 | 8524 | 1200 | 71290 |
| | Dépensé | 4120 | 702 | 1200 | 51254 |
| | Engagé | 3050 | 0 | 0 | 20036 |
| | Pourcent | 44.8% | 8.2% | 100% | 99% |
| TYREX | Alloué | 29000 | 10008 | 5100 | 68400 |
| | Dépensé | 7568 | 209 | 4480 | 0 |
| | Engagé | 1550 | 4406 | 0 | 68400 |
| | Pourcent | 31.4% | 46.1% | 87.8% | 100% |

# Auto-évaluation (AFOM)

## Atouts

- Expertise internationalement reconnue pour chaque partenaire
- Equipes à la pointe et en avance sur des sujets stratégiques
- Publications dans les meilleures conférences
- Excellente entente entre les partenaires

## Faiblesses

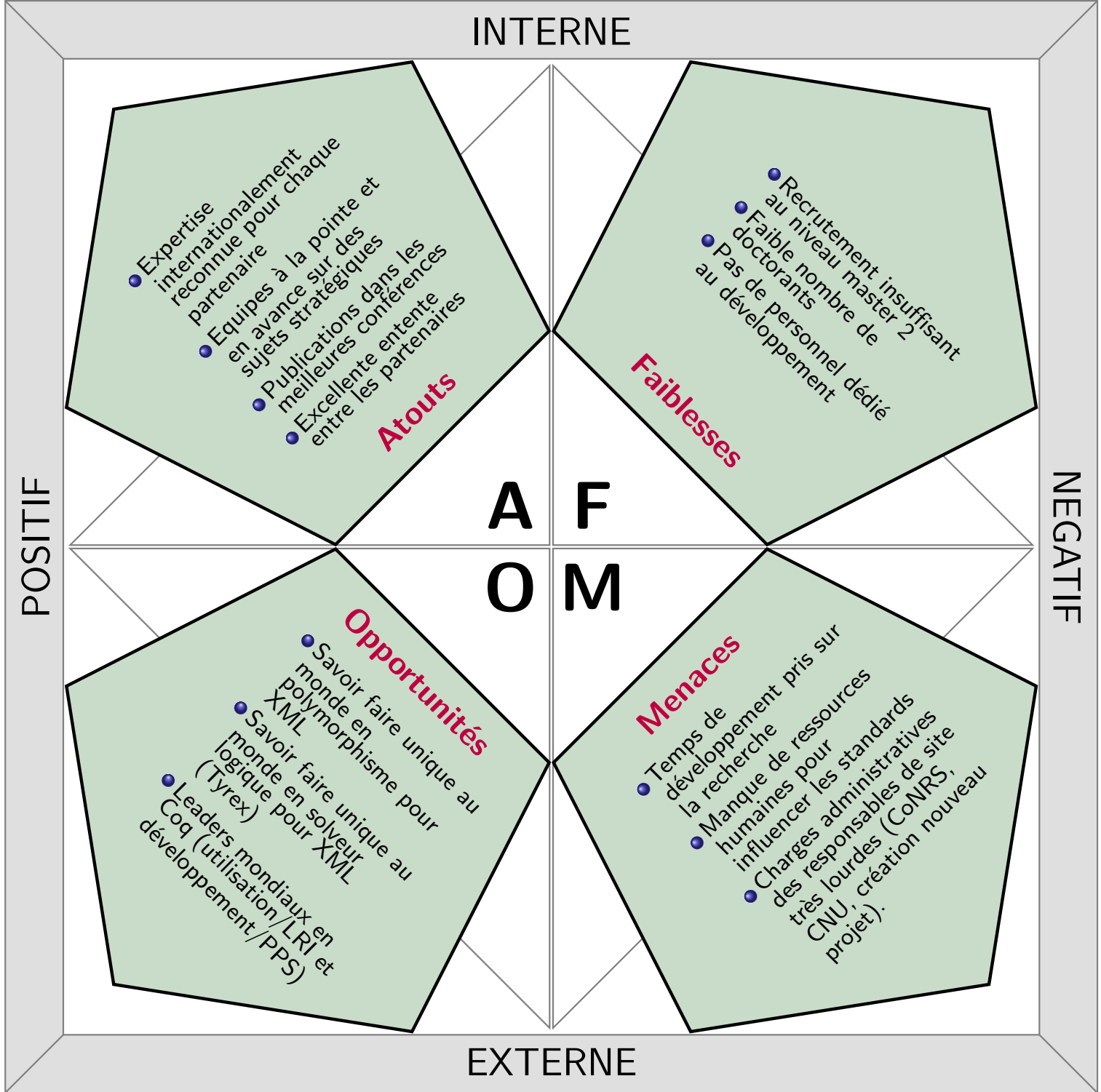
- Recrutement insuffisant au niveau master 2
- Faible nombre de doctorants
- Pas de personnel dédié au développement

## Opportunités

- Savoir faire unique au monde en polymorphisme pour XML
- Savoir faire unique au monde en solveur logique pour XML (Tyrex)
- Leaders mondiaux en Coq (utilisation/LRI et développement/PPS)

## Menaces

- Temps de développement pris sur la recherche
- Manque de ressources humaines pour influencer les standards
- Charges administratives des responsables de site très lourdes (CoNRS, CNU, création nouveau projet).


INTERNE
POSITIF
NEGATIF
EXTERNE
A F
O M
Atouts
Expertise internationalement reconnue pour chaque partenaire
Equipes à la pointe et en avance sur des sujets stratégiques
Publications dans les meilleures conférences
Excellente entente entre les partenaires
Faiblesses
Recrutement insuffisant au niveau master 2
Faible nombre de doctorants
Pas de personnel dédié au développement
Opportunités
Savoir faire unique au monde en polymorphisme pour XML
Savoir faire unique au monde en solveur logique pour XML (Tyrex)
Leaders mondiaux en Coq (utilisation/LRI et développement/PPS)
Menaces
Temps de développement pris sur la recherche
Manque de ressources humaines pour influencer les standards
Charges administratives très lourdes (CoNRS, des responsables de site CNU, création nouveau projet).