Yannick Perret – CNRS

Système de cache pour l'extraction HAL

Laboratoire d'InfoRmatique en Image et Systèmes d'information

LIRIS UMR 5205 CNRS/INSA de Lyon/Université Claude Bernard Lyon 1/Université Lumière Lyon 2/Ecole Centrale de Lyon

http://liris.cnrs.fr

Publications LIRIS : contexte

Gestionnaire de publications interne (~2006)

- **Intégré au SI du laboratoire (liens internes…)**
- **Outils spécifiques (bilans, recherches, tris…)**
- **Vues labo / équipes / membres + publications externes**
- **Intégration dans des CMS tiers (dokuwiki…)**
- **Les membres saisissent leurs publications**

→ HAL existait mais n'offrait pas la visibilité et les fonctionnalités actuelles

Publications LIRIS : bascule HAL

2014 → contexte différent :

- **SI / CMS LIRIS vieillissant**
- **Gestionnaire de publications fortement lié au SI**
- **Outils d'import vers HAL de l'existant**
- **HAL devient « visible » (recommandé/imposé par certaines tutelles)**
  - **D'ailleurs cela imposait la double saisie à ces membres**
- **HAL gère des éléments nécessaires :**
  - **IdHAL → identifier un membre du laboratoire**
  - **Équipes → nécessaire pour le LIRIS (et d'autres !)**

De plus fin du quadriénal en cours pour le LIRIS
→ changements « plus simples »

Gestionnaire LIRIS / HAL

Nécessité d'avoir notre propre « vue » des publications :

- **Intégration au SI (liens, filtres spécifiques…)**
- **Mise en forme / données affichées spécifiques (en particulier durant la transition, avec affichage mixte)**
- **Certaines métadonnées internes non gérables dans HAL**

De plus certains traitements sont nécessaires pour gérer des cas non pris en compte par HAL (ou certains déposants) :

- **Cas d'équipes «spéciales » (non séparables par l'API)**
- **Thèses difficiles à gérer (délais, absences, doublons…)**

Nécessité d'un cache local

De multiples raisons :

- **Ne pas surcharger HAL à chaque requête sur notre site**
- **Gagner en réactivité (moins de requêtes intermédiaires)**
- **Conserver un affichage indépendant de HAL :**
  - **En cas de ralentissement**
  - **En cas de maintenance ou de panne**
- **Gain en efficacité :**
  - **Publications stockées localement sous forme « traitée »**
  - **Métadonnées internes ajoutées en cache**
  - **Traitements à vitesse du disque (et non du réseau)**

De plus les publications ne changent pas en permanence
→ compatible avec une vision « cache »

Structure du cache

Données internes :

- **Liste des membres (id interne ↔ id HAL)**
- **Liste des équipes (id interne ↔ id de structure)**
- **Liste ~~des~~ du laboratoire :) (identifiant ↔ id de structure)**
- **Liste de données spécifiques (métadonnées internes)**
- **Certaines préférences des outils**

Cache : chaque publication stockée en cache séparément

Listes : liste d'identifiants de publications dans le cache, pour chaques membres/équipes/laboratoires

Algorithme : principe

Via **l'API HAL, lister les identifiants de publications qui sont rattachées au LIRIS et modifiées depuis XXXX (date de dernière mise à jour du cache).**
→ ces publications sont à re-charger / analyser en cache.
⇒ seules celles-ci nécessitent un accès « complet » sur HAL

Note : ceci ne gère pas les publications qui « sortent » du LIRIS et restent en cache, inutiles.
→ plusieurs solutions :

- **Ne rien faire**
- **Lister toutes les publis et supprimer**
- **Purger régulièrement (*reset* du cache)**
- **Péremption des entrées du cache**
- **Supprimer les entrées non présentes en listes**

Algorithme : publications à traiter

- **API : liste des pub-id [affiliation LIRIS + modifié depuis XX]**
  - **∀id → inséré dans liste de MàJ**
- **Pour chaque membre (IdHAL) :**
  - **API : liste des pub-id [IdHAL + modifié depuis XX]**
    - **∀id → inséré dans liste de MàJ**

→ liste des publications nécessitant récupération

Algorithme : récupération+traitement

- Pour chaque pub-id dans la liste de MàJ :
  - API : récupération des données
  - Analyse et construction d'un objet « publication »
  - Modification de l'objet en fonction des données internes
  - Stockage en cache sous forme d'objet sérialisé

+ sauvegarder la date de dernière mise à jour

Algorithme : aiguillage

Cette phase est indépendante des précédentes

- **Pour chaque publication dans le cache :**
  - **Lecture / dé-sérialisation**
  - **Aiguillage de la publication (labo / équipes / membres)**
    - **Ajout dans les listes dédiées**
    - **Analyses de cohérences, de problèmes…**

Suppression des entrées inutiles :

- **∀ entrée présente en cache**
  - **Si pas aiguillé (labo ou équipe ou …)→ supprimer**

- **Mise à jour des listes, pages de statistiques…**

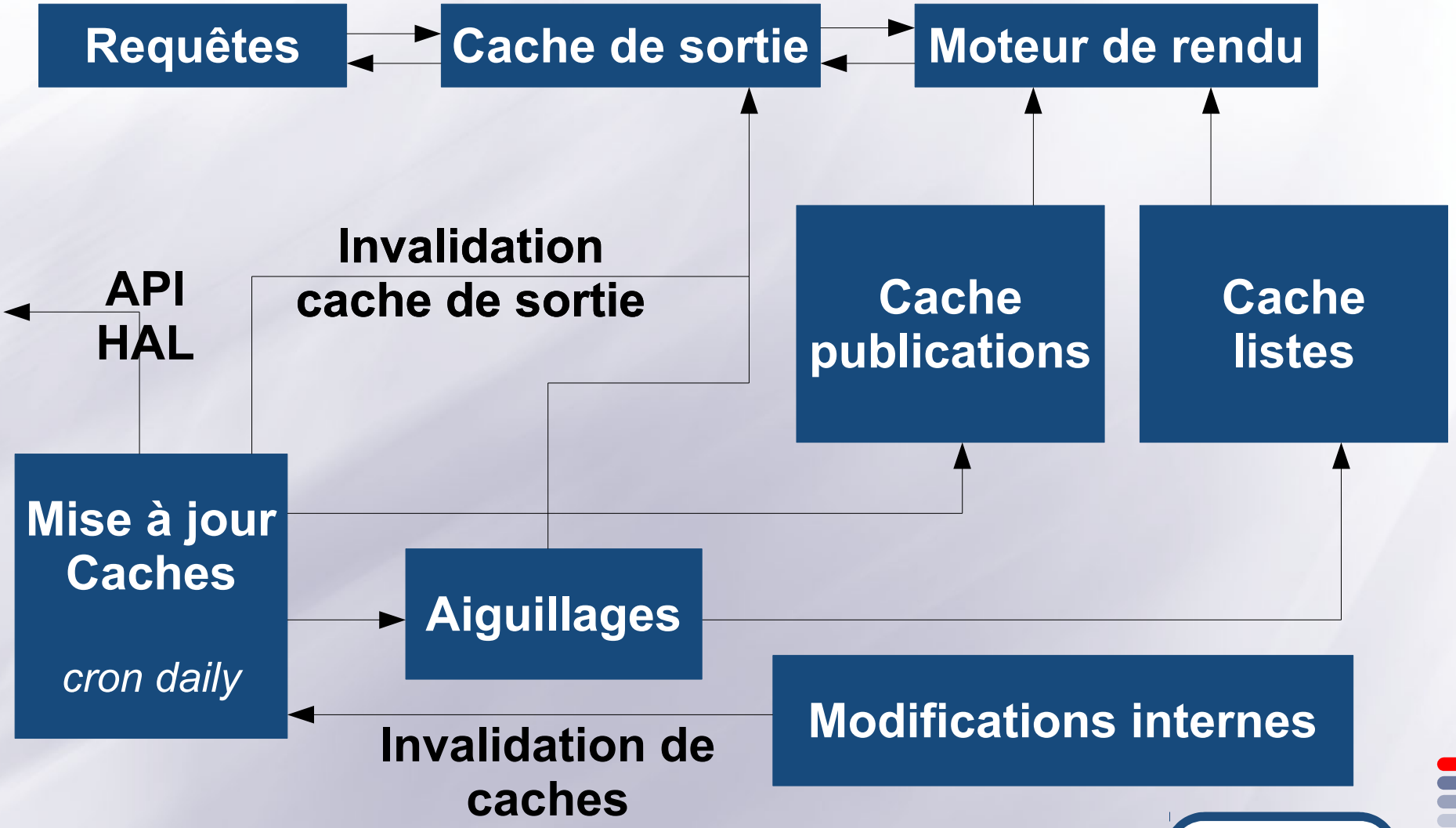
Invalidation de caches

Nécessaire dans plusieurs cas :

- **Évolution de la structure de l'objet publication (re-parse)**
- **Évolution données membres/équipes…**
  - **Lister les publications concernées + invalider**
- **Problème de stockage (erreur, écrasement…)**
- **Choix techniques**

Également nécessaire car les fusions d'auteurs ne sont pas vues comme des modifications de publications
→ nécessite une analyse externe et une invalidation spécifique (bug report / demande de fonctionnalité ouvert chez HAL)
⇒ choix d'une péremption (+6j)

Flux fonctionnel



Algorithme : performances

LIRIS : 3560 publications + 1300 publications externes
En moyenne ~40 publications à mettre à jour

- **Créer la liste des publications à mettre à jour : ~10s**
- **Récupérer+analyser+stocker ces 40 pubs : ~50s**
- **Lire+Aiguiller toutes les pubs en cache : ~15s**

- **Taille moyenne sur disque : 9ko par publication**
- **Taille totale cache+listes : ~40Mo**

Récupération « en une fois » sur HAL
→ Internal Server Error

Algorithme : affichage

- API WEB (type, nom, options, filtres…)
- Lecture de la liste associée. Pour chaque publication :
  - Lire + dé-sérialiser la publication
  - Appliquer les filtres
- Appliquer les tris
- Générer le rendu

+ présence d'un cache de sortie

Rendu 3560 publications (classées+triées) : ~7s
Rendu avec cache d'affichage : ~0.9s

Divers : fonctionnalités du rendu

- **Filtres : par membre, équipe, dates, types, mots-clés**
- **Formats : HTML, XML, JSON, CSV, HTML-raw**
- **Statistiques : CSV, lié aux demandes secrétariat**
- **Tris : année+type / type+année / type / …**
- **Membres : publications internes, externes, toutes**
- **Personnalisations du rendu**
- **HTML : CSS couvrant chaque élément de sortie**
- **Configuration : préférences par défaut (tris, format d'affichage, éléments affichés…)**

Plugins d'intégration dans nos fermes de CMS/wiki :

- **Dokuwiki**
- **Drupal**

Divers : fonctionnalités d'analyse

Lors de l'analyse / aiguillage diverses analyses sont effectuées sur les publications :

- **Mauvaise affiliation (LIRIS mais pas équipe)**
- **Absence d'IdHAL (affilié LIRIS mais sans IdHAL)**
  - **Recherche membres potentiellement concernés**
- **Détection de doublons possibles (thèses, principalement)**
- **Cohérence de dates (présence membres)**
- **Données manquantes sur HAL**

Exemple

HAL : hal-01460709.

HDR, thèses (3)

Thèses (3)

- Magali Ollagnier-Beldame (2017). « Interaction Traces and cognitive process in a joint activity ». HAL : hal-01464161.
- Maxime Gasse (2017). « Probabilistic Graphical Model Structure Learning: Application to Multi-Label Classification ». HAL : tel-01442613.
- Yolanda Sanchez-Dehesa (2017). « RÆvol : un modèle de génétique digitale pour étudier l'évolution des réseaux de régulation génétiques ». HAL : hal-01462010.

Éditions scientifique d'ouvrages (livres, chapitres, colloques, congrès, n° spéciaux) (3)

- Jakob Bardram, Morten Esbensen & Aurélien Tabard (2017). « Activity-Based Collaboration for Interactive Spaces ». Collaboration Meets Interactive Spaces, pp. 233-257. doi : 10.1007/978-3-319-45853-3_11. HAL : hal-01436498.
- Kanishk Chaturvedi, Carl Stephen Smyth, Gilles Gesquière, Tatjana Kutzner & Thomas H. Kolbe (2017). « Managing Vers Chapitre d'ouvrage, audience internationale Next Generation of CityGML ». Lecture Notes in Geoinformation and Cartography, Abdul-Rahman, Alias, Springer, pp. 191-206. doi : 10.1007/978-3-319-25691-7_11. HAL :
- Florence Zara & Olivier Dupuis (2017). « Chap 15: Uterus – Biomechanical modeling of uterus. Application to a childbirth simulation ». Biomechanics of Living Organs: Hyperelastic Constitutive Laws for Finite Element Modeling, Yohan Payan, Jacques Ohayon, Elsevier. HAL : hal-01486956.

2016 (312)

Revues (95)

Revues internationales avec comité de lecture (89)

- Zohra Saoud, Noura Faci, Zakaria Maamar & Djamal Benslimane (2016). « A Fuzzy-based Credibility Model to Assess Web Services Trust under Uncertainty ». Journal of Systems and Software, . HAL : hal-01207317.
- Djamila Boukredera, Ramdane Maamri & Samir Aknine (2016). « Stochastic Petri Net-Based Modeling and Formal Analysis of Fault Tolerant Contract Net Protocol ». Web Intelligence and Agent Systems: An International Journal, vol. 14, n°3, pp. 245-271. HAL : hal-01254503.
- Djamal Benslimane, Quan Z. Sheng, Mahmoud Barhamgi & Henri Prade (2016). « The Uncertain Web: Concepts, Challenges, and Current Solutions ». ACM Transactions on Internet Technology, . HAL : hal-01255223.
- Abdelhamid Malki, Djamal Benslimane, Sidi-Mohamed Benslimane, Mahmoud Barhamgi, Mimoun Malki, Parisa Ghodous & Khalil Drira (2016). « Data Services with uncertain and correlated semantics ». World Wide Web, vol. 19, n°1, pp. 157-175. HAL : hal-01208088.
- Marthe Bonamy, Nicolas Bousquet & Stéphan Thomassé (2016). « The Erdös--Hajnal Conjecture for Long Holes and Antiholes ». Siam Journal on Discrete Mathematics, vol. 30, n°2, pp. 1159-1164. doi : 10.1137/140981745. HAL : lirmm-01347304.
- Antoine Boutet, Davide Frey, Rachid Guerraoui, Arnaud Jégou & Anne-Marie Kermarrec (2016). « Privacy-Preserving Distributed Collaborative Filtering ». Computing, . HAL : hal-01251314.
- Victor Charpenay, Elod Egyed-Zsigmond & Harald Kosch (2016). « Knowledge-driven reverse geo-tagging for annotated images ». Revue des Sciences et Technologies de l'Information - Série Document Numérique, vol. 19, pp. 83-102. doi : 10.3166/DN.19.1.83-102. HAL : hal-01343362.
- Angela Bonifati, Radu Ciucanu & Slawomir Staworko (2016). « Learning Join Queries from User Examples ». ACM Transactions on Database Systems, vol. 40, n°4, 24:1-24:38. HAL : hal-01187986.