Stage de fin d'études pour le diplôme collégial

Centre de Données astronomiques de Strasbourg

Projet d'homogénéisation des logs des différents services du CDS

Ordre du jour

- Introduction du stage
- Explication du projet
- Présentation du travail accompli
- Problèmes rencontrés
- Travail restant à accomplir
- Ce que je retire du stage

Introduction du stage

- Accueil au CDS
- Lieu de travail → Bibliothèque
- Le CDS c'est :
  - À peu près 30 employés en tout
  - Le tiers en informatique donc environ 10 personnes
  - Présentement une douzaine de stagiaires en informatique
- 3 services principaux

3 services principaux

SIMBAD Astronomical Database

Queries	Documentation	Information
basic search	User's guide	Presentation
by identifier		
by coordinates		Acknowledgment
by criteria	Query by urls	
reference query	Nomenclature Dictionary	
scripts	Object types	
TAP queries	List of journals	
options	Measurement description	
	Spectral type coding	Release:
Display all user annotations	User annotations documentation	SIMBAD4 1.222 - 25-Apr-2014

Overview

Aladin lite is a lightweight version of the Aladin tool, running in the browser and geared towards simple visualization of a sky region.

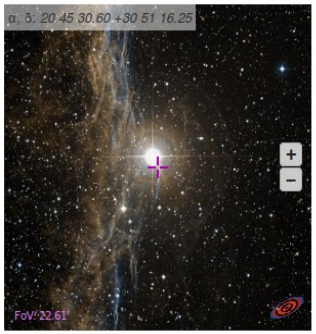
It allows one to visualize image surveys (JPEG multi-resolution HEALPix all-sky surveys) and superimpose tabular (VOTable) and footprints (STC-S) data.

Aladin lite is powered by the HTML5 canvas technology, currently supported by any modern browser

Aladin lite is easily embeddable on any web page and can also be controlled through a Javacript API.

It is dedicated to replace the Aladin Java applet technology in the medium term.

The panel on the right hand is not a regular image. It is actually Aladin Lite running as an embedded widget. You might try to zoom in and out using your mouse wheel, or pan the view to move around.



VizieR Service

new **The CMC15 and IGSL3 catalogues are available in VizieR.**

Find catalogs among 12314 available

Clear | Find... | Expand search

Catalog, author's name, word(s) from title, description, etc. e.g.: AGN, Veron, I/239, or bibcodes...

Search for catalogs by column descriptions *(UCD)*

Search for catalogs containing additional data

Wavelength	Mission	Astronomy
Radio	AKARI	Abundances
IR	ANS	Ages
optical	ASCA	AGN
UV	BeppoSAX	Associations
EUV	CGRO	Atomic_Data
X-ray	Chandra	Binaries:cataclysmic
Gamma-ray	COBE	Binaries:eclipsing

Search by Position across 12950 tables

Target Name (resolved by Sesame) or **Position**: | Target dimension:

Clear | J2000 | 2 | arcmin | Go!

Radius | Box size

More about VizieR | Find Catalogs

Explication du projet

- Situation actuelle à l'observatoire
  - Les logs sont éparpillés au niveau des différents services
  - Faire des statistiques peut prendre jusqu'à 3 jours
  - Les logs ne sont pas du tout sous le même format

Explication du projet

- Différents objectifs du stage
  - Définir un format pivot qui serait utilisé par n'importe quel type de log du CDS
  - Stocker tout ces logs sur une base de données
  - Faire des tests de performances sur la base de données pour s'assurer du bon choix
  - Développer un outil de génération de statistiques qui ferait appel à la base de données

Présentation du travail accompli

- Programmes de traduction vers le format pivot
  - Définition claire du format pivot

identificateur	adresse_ip	date	service	query-string	method
0000001	130.79.30.10	2010/09/10 23:12:03	Simbad	id=M31 coord=5.93.34 magn=45	Sim-id
0000002	10.69.39.129	2010/09/10 23:12:04	Aladin	-	Allskyimage

    - Format JSON

  - Écriture de 3 programmes de traduction
    - Simbad HTTP
    - VizieR
    - Aladin

Présentation du travail accompli

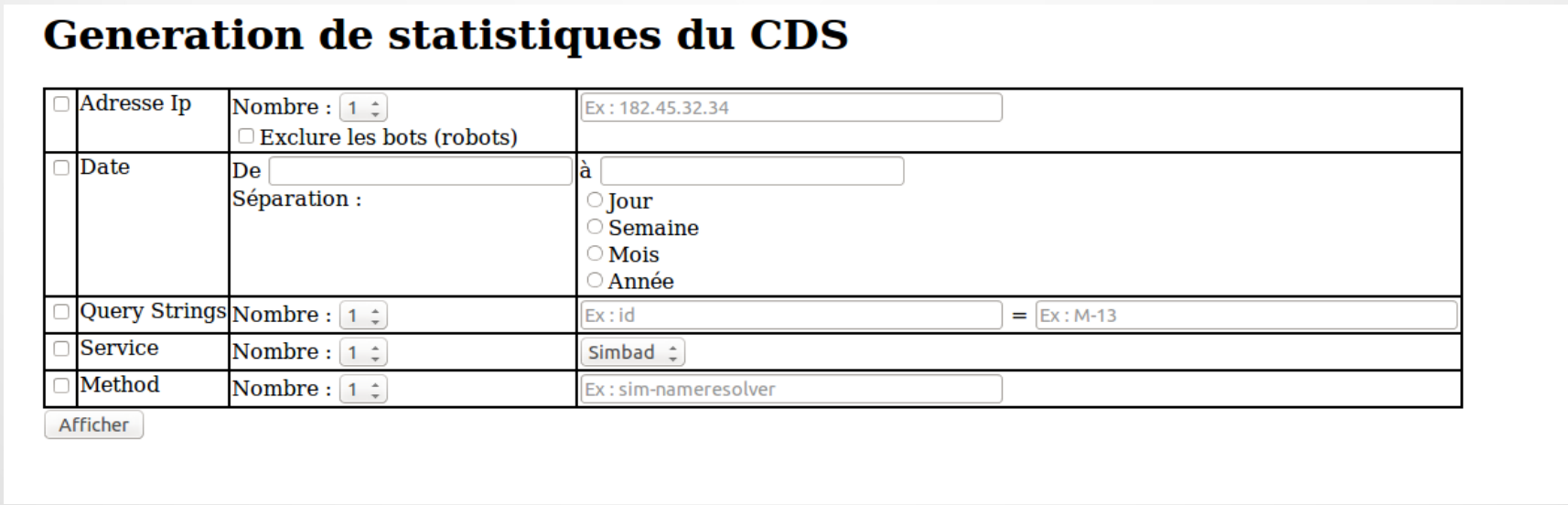
- Programme d'importation vers la base de données MongoDB
  - Reçoit les fichiers JSON générés par les différents programmes de traduction et les insère dans la base d données.
  - Système de vérification : On regarde si on a pas déjà inclus le fichier qu'on tente d'importer à l'aide d'un del

Présentation du travail accompli

- Tests de performances sur les différents aspects de MongoDB
  - Tests « grandeur-nature » pour tester l'outil dans un contexte d'utilisation réelle
  - Tests sur la mémoire utilisée lors de différents états de MongoDB (requête, pendant l'importation, état mort, etc.)
  - Regard sur les temps d'importation de différentes années et le volume de celle-ci

Présentation du travail accompli

- Interface de génération de statistiques (Pas encore finie)
  - Pour l'instant seulement codée en HTML, il faut ajoute le PHP et trouver un moyen de communiquer avec le serveur MongoDB.

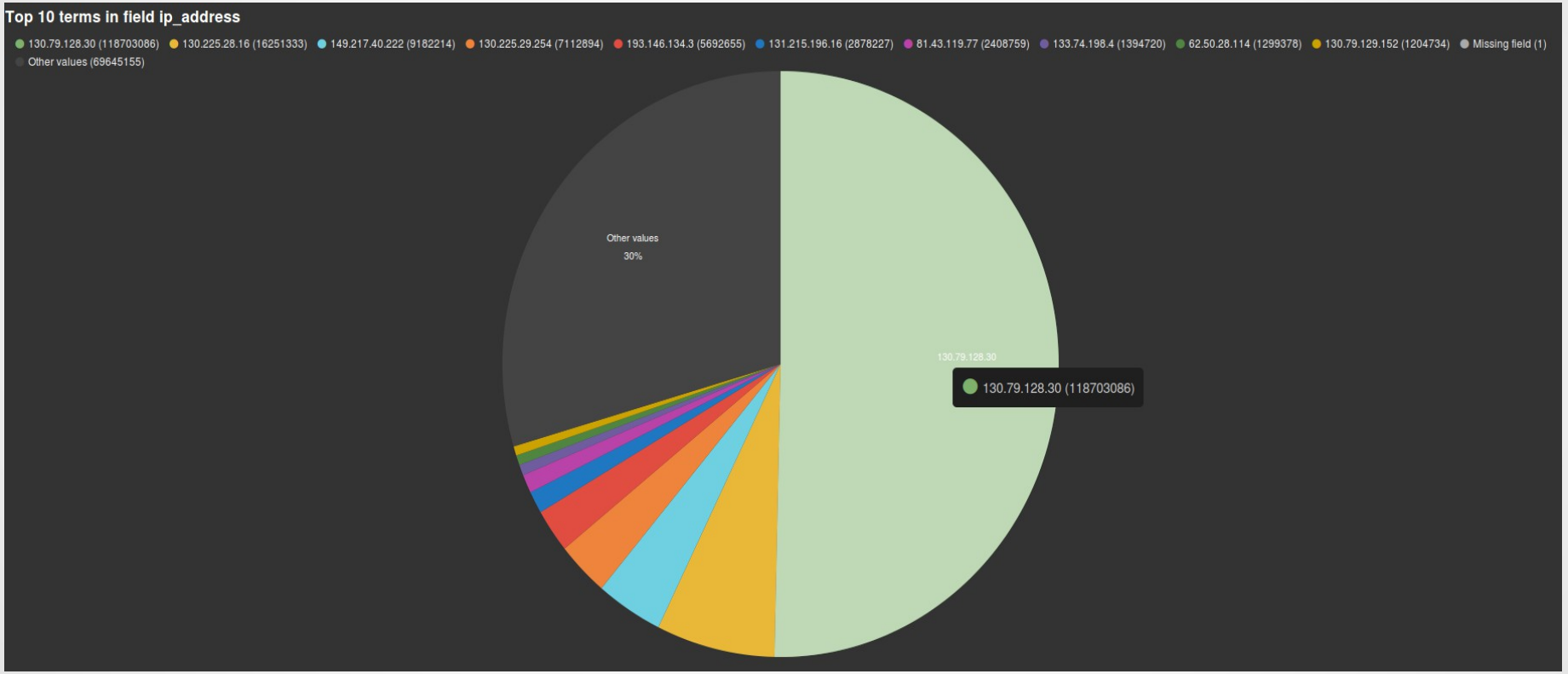


Generation de statistiques du CDS

☐	Adresse Ip	Nombre : 1 ☐ Exclure les bots (robots)	Ex : 182.45.32.34
☐	Date	De Séparation :	à ○ Jour ○ Semaine ○ Mois ○ Année
☐	Query Strings	Nombre : 1	Ex : id = Ex : M-13
☐	Service	Nombre : 1	Simbad
☐	Method	Nombre : 1	Ex : sim-nameresolver

Afficher

Kibana


Top 10 terms in field ip_address
130.79.128.30 (118703086)
130.225.28.16 (16251333)
149.217.40.222 (9182214)
130.225.29.254 (7112894)
193.146.134.3 (5692655)
131.215.196.16 (2878227)
81.43.119.77 (2408759)
133.74.198.4 (1394720)
62.50.28.114 (1299378)
130.79.129.152 (1204734)
Missing field (1)
Other values (69645155)
Other values
30%
130.79.128.30
130.79.128.30 (118703086)

Problèmes rencontrés

- Temps de réponse de MongoDB devenu lent avec toutes les entrées
  - Après avoir inséré 400 millions d'entrées, le temps de requête est devenu très lent pour MongoDB (18 minutes)
  - Il faut trouver un moyen d'optimiser MongoDB ou bien d'indexer les données

Problèmes rencontrés

- Indexation des données avec ElasticSearch
  - Dans l'optique de vouloir obtenir des réponses de requêtes plus rapides, j'ai tenté d'indexer les données avec un outil spécialisé pour ce genre de chose : ElasticSearch
  - Incapable de faire une indexation complète sans rencontrer un problème de Java Heap Space, même paramétré avec plus de mémoire

Problèmes rencontrés

- Connexion de PHP à MongoDB
  - Malgré les exemples et les tutoriels vus sur le web, incapable de me connecter à MongoDB
  - Driver non-fonctionnel ?
  - <?php

    $m = new MongoClient(); // connect

    $db = $m->selectDB("example");

    ?>

Travail restant à accomplir

- L'objectif initial (de rendre tout les logs sont un même format et de les stocker sur le même endroit) à été atteint mais il reste des choses à faire :
  - Optimiser MongoDB ou indexer les données d'une façon ou d'une autre
  - Finir l'outil de statistiques en PHP

Ce que je retire du stage

- Niveau technique
  - Travail de précision avec les chaînes de caractères
  - Découverte de MongoDB
  - Découverte de ElasticSearch (et Kibana)
  - Environnement Linux

Ce que je retire du stage

- Niveau personnel
  - Appris à travailler en équipe et participation à des réunions avec les superviseurs
  - Participation et présentation à la réunion Infusion devant une douzaine d'informaticiens

Merci