Prototypage d'une base de données bibliographiques en interférométrie optique

Esther TAILLIFET

Magistère $2^{ème}$ année - Projet informatique 2009-2010

Table des matières

1 Introduction

La deuxième année du Magistère de Physique de l'université Joseph Fourier Grenoble I propose un module intitulé "Projet informatique". Il s'agit d'un court stage d'informatique à effectuer en binôme au sein de l'un laboratoire de la ville. Le contenu du projet à l'avantage et à la fois l'inconvénient d'être totalement libre, les seules contraintes étant la durée fixée à quarante heures et l'usage de l'outil informatique. C'est donc sans trop savoir ce que je cherchais qu'à l'automne, j'ai proposé mes petites mains peu qualifiées mais volontaires aux chercheurs du Laboratoire d'astrophysique de Grenoble (Laog). Fabien Malbet est l'un de ceux à avoir répondu à mon appel, voyant là l'occasion d'amorcer l'un de ses projets : la construction d'une base de données bibliographiques en interférométrie optique.

2 Présentation du projet

2.1 Un projet utile

L'interférométrie connait un essor important ces dernières années notamment par le nombre d'instruments croissant et un intérêt récent pour cette technique. Les publications sont donc de plus en plus nombreuses. Depuis plusieurs années Peter Lawson (chercheur de l'Institute of technology, California) met à jour un site web répertoriant sous forme de liste chronologique ces publications : Olbin (Optical Long Baseline Interferometry Newsletter) consultable à l'adresse suivante : http ://olbin.jpl.nasa.gov/. C'est un travail qui lui prend du temps et qui est malheureusement difficilement exploitable. C'est pour cette raison que Fabien Malbet souhaitait créer une base de données bibliographiques répertoriant les publications en interférométrie optique qui pourrait remplacer cette liste. Le contenu des articles ainsi qu'un certain nombre d'informations utiles seraient ainsi facilement consultables par la communauté scientifique. L'objectif étant également d'éviter à Peter Lawson ce travail fastidieux de mise à jour en donnant la possibilité aux chercheurs d'entrer eux même leurs nouvelles publications dans la base de données. Pour que cette base soit la mieux pensée possible, Fabien s'est entouré d'informaticiens au sein du Laog, dont Guillaume Mella, et m'a proposé de me joindre a son projet à l'occasion de ce stage.

2.2 Les objectifs du stage

Guillaume avait déjà réfléchi à la façon dont nous allions construire cette base de données d'un point de vue technique et était donc prêt pour nous y aider. L'idée de Fabien était de créer un premier prototype à partir de la liste bibliographique disponible sur Olbin. Il ne s'agissait pas de refaire le travail bibliographique déjà fait par Peter Lawson de manière certainement très complète, mais de permettre son exploitation. L'objectif global du stage était donc d'amorcer la construction de la base de données. Plus précisément, pour la quarantaine d'heures qui m'était données, nous avions fixé les objectifs suivant :

– définir les informations à faire figurer dans la future base de données

– définir les paramètres de sorties, c'est à dire les paramètres sur lesquels nous souhaitons pouvoir l'interroger.
– insérer dans la base de données toute les références présentes sur olbin.
– modifier le code PHP de la page Olbin existante pour faire apparaître le contenu de la base de données à la place de la liste bibliographique.
– interroger la base de données de façon simple sur certains de ces paramètres et de créer une page web présentant les résultats obtenus.
– créer la page web qui permettra d'introduire de nouvelles entrées bibliographiques dans la base.

L'objectif pour moi était donc de comprendre l'arborescence d'une base de données et de m'initier à plusieurs types de langages informatique (HTML, PHP et MySQL).

2.3 Les raisons de ma participation à ce projet

Avant d'accepter de participer à ce projet il m'a fallut l'accord de Laurent Derôme pour le faire seule car l'une des rares consignes pour valider le stage dans le cadre du Magistère était d'effectuer le projet en binôme.

Plusieurs choses m'ont poussées à choisir ce stage. Tout d'abord, j'ai trouvé le projet utile. Cette base de données allait être un outil de travail pour les chercheurs dans les mois et les années avenir. J'étais enthousiaste à l'idée de participer à l'initiation d'un projet comme celui-ci et malgré tout être en mesure de présenter des résultats après seulement quarante heures de stage. Le projet m'a semblé adaptable à mes besoins et à mon niveau en informatique. Je n'avais jamais manipuler une base de données ni construit de page web, et j'ai pensé que se serait une bonne occasion de m'initier. J'ai préféré un projet comme celui-ci à un autre en Matlab ou en C^{++} où j'avais déjà quelques notions. J'ai supposé que l'objectif de ce stage de Magistère au contenu très libre, était de nous laisser prendre l'initiative de découvrir des facettes de l'informatique que nous n'apprenons pas sur les bancs de l'université dans nos filières. La dernière chose ayant influencé mon choix est tout simplement la rencontre avec Fabien et Guillaume. J'ai vu que j'allais pouvoir apporter ma pierre à l'édifice et être à l'aise de leur poser mes questions. Aussi secondaire que cela puisse paraître, je pense qu'une bonne entente et une bonne compréhension au sein d'un groupe de travail sont essentielles pour réussir un stage constructif autant pour le stagiaire que pour son maitre de stage.

3 Construction du prototype de base de données

Pour travailler ensemble, la création d'un cahier des charges était essentielle afin de pour partager notre travail et créer des pense-bêtes pour ne rien oublier. C'est sur le serveur jmmc qu'a été loger ce cahier des charges à l'adresse suivante : http ://www-laog.obs.ujf-grenoble.fr/twiki/bin/view/Laog/GRIL/Informatique/JmmcInterferometryDB . Je me suis également efforcée de mettre à jour ce que j'ai appelé mon "Bilan journalier" afin de me souvenir de se que j'avais fait pendant chacune de mes venues au Laog.

3.1 Le choix du contenu

La première étape dans la construction d'une base de données est de se mettre d'accord sur son contenu. Pour savoir quelles informations allaient être utiles, il a fallu se poser la question de l'usage qui allait être fait de la base. C'est se que nous nous sommes efforcés de faire tout au long de la construction du prototype.

3.1.1 Les entrées

Pour construire l'arborescence de notre base de données nous nous sommes servit des outils à notre disposition. Le site Olbin de Peter Lawson servant de point de départ et ADS étant notre principale source s'informations. ADS (Astrophysics Data System) est une base de données accessible en ligne (http ://cdsads.u-strasbg.fr/) répertoriant toutes les publications en Astronomie et Astrophysique. Chaque publication est classée à l'aide d'un code bibliographique (ou bibcode [1]) qui donne accès à son abstract. Chaque bibcode donne également un lien vers arXiv où sont stockées les publications scientifiques sous un format électronique. Il était donc utile de faire figurer les bibcodes des articles dans notre base de données et de créer un lien vers ADS permettant d'accéder au contenu de l'article. Outre le bibcode de l'article, un certain nombre de "données principales" devaient être répertoriées :

- Le titre complet de la publication
- Les auteurs, leurs nationalités et leurs affiliations
- La date de publication
- Le journal dans lequel l'article a été publié

Ces données se trouvaient déjà sur Olbin pour chacune des références ou étaient facilement récupérables sur ADS.

Nous avons également jugé intéressant d'introduire un certain nombre de données complémentaires à celles déjà présentes sur Olbin comme :

- Les mots clés
- Le type de papier : observations, théorie, instrumentation.
- Les éventuels résultats
- Les théories, les prédictions
- Les autres publications en lien direct
- Le site d'observation (VLTI, IOTA, PTI, CHARA,...)
- L'instrument (AMBER, MIDI, PRIMA, MIRC, VEGA,...)
- Les domaines de longueur d'ondes (visible, NIR, MIR,...)
- Le type de mesure effectuée (V, phase, clôture de phase, images, nulling,...)
- Le type d'objet observés

Il allait donc être nécessaire de prendre chaque publication indépendamment et d'entrer manuellement chacune de ces "métadonnées" en parcourant le contenu des articles.

[1] Le bibcode est un code bibliographique à dix-neuf caractères utilisé par ADS. Il est construit à partir de l'année de publication, du nom, numéro et page du journal dans lequel il a été publié et de l'initiale de l'auteur principal. Par exemple : 2007AA...464....1P est le bibcode d'un article publié en 2007 dans en première page du $464^{ème}$ numéro du célèbre journal *Astronomy and Astrophysics*. Le premier auteur de l'article, R.G Petrov prête son initiale au bicode.

3.1.2 Les sorties

Nous avons jugé intéressant de choisir de consulter la liste des publications par années, par auteurs ou par instrument par exemple. Nous pouvions alors imaginer aussi pouvoir effectuer des statistiques sur le nombre d'articles publiés par année, ou par auteur, ou encore savoir le nombre de références contenues dans la base de données, le nombre d'objets différents référencés, ou même des requêtes combinées. C'était là l'un des objectifs de cette base de données, donner une "plus value" au travail de Peter Lawson.

3.2 La création du prototype

3.2.1 Listes de bibcodes

Pour pouvoir remplir la base de données, il nous fallait tout d'abord faire la liste de tout les bibcodes correspondants aux publications référencées sur Olbin. C'est moi qui me suis occupée de cette tâche. Il a fallu pour chaque article retrouver le bibcode correspondant en cherchant chaque titre sur ADS. Ce n'est pas un travail difficile, mais il faut être concentré afin de ne pas oublier un article ni de se tromper de publication en copiant le bibcode. Pour les deux années les plus récentes, Peter Lawson avait commencer à mettre un lien vers ADS pour presque tous les articles. Or lorsque nous sommes sur la page ADS d'un article référencé, l'url se termine par le bibcode de l'article. Il était alors facile pour moi en regardant le code PHP de la page Olbin et en utilisant astucieusement emacs, de trier les informations afin de laisser juste les dix-neufs derniers caractères du lien correspondants au bibcode de l'article. Il restait tout de même quelques articles à chercher manuellement. Pour les années précédentes en revanche, il m'a fallut chercher l'intégralité des articles manuellement sur ADS sachant que le nombre de publications allait en décroissant.

Pour éviter de faire des erreurs, je vérifiais après avoir finit une année, que le nombre de bibcodes était identique au nombre de publications sur Olbin. Il a également fallu vérifier de manière automatisé (par comparaison de fichiers) qu'aucun bibcodes n'apparaissaient deux fois dans les listes. Évidement nous avons détecter un certain nombre d'erreurs liées au fait que ces listes ont étés crées à la main. Quand ce n'était pas moi qui répétait un bibcode par erreur alors c'était Peter Lawson qui avait fait figurer deux fois un article sur Olbin.

3.2.2 Données principales

Une fois la liste des bibcodes prête, nous étions capables de remplir la base de données avec les données principales disponible sur ADS à partir des bibcodes. C'est Guillaume qui s'est chargé d'écrire le script permettant de parcourir les fichiers de la forme `papers-****.txt` contenants les bibcodes et d'aller récupérer les informations concernant ces publications sur ADS. Voici l'ensemble des commandes à faire à partir du terminal pour insérer les données principales des publications dans la base de donnée :

```
1.  ssh jmmc.fr
2.  cd public_html/bibdb/2_feedDB/
```

```
3.  wget http://www-laog.obs.ujf-grenoble.fr/twiki/pub/Laog/GRIL/
    Informatique/JmmcInterferometryDB/papers-2009.txt
4.  ./genADS.sh
5.  xsltproc AdsToSql.xsl papers-2009.txt.xml > papers-2009.txt.xml.sql
6.  read -s PASSWD
7.  mysql -u bib --password=$PASSWD bib < papers-2009.txt.xml.sql
```

If faut tout d'abord se connecter au serveur web ("ssh" signifie "secured shell") c'est à dire se connecter à "la machine jmmc" où se trouvent tout les fichiers concernant la base de données. La commande 2 permet de se placer dans le bon dossier. La troisième ligne de commande permet d'aller chercher dans ce dossier, le fichier contenant les bibcodes à introduire dans la base (en l'occurrence ceux de 2009). La commande suivante lance le script de récupération des données sur ADS. Un fichier `xml` contenant les données ADS est créé. La commande 5 permet de transformer les données ADS en requêtes `Sql`(Structured Query Language), qui est un langage informatique standardisé permettant d'interroger et de manipuler des bases de données. L'insertion proprement dite des données dans la base se fait à l'aide des deux dernières commandes.

Lors de cette étape nous nous somme aperçu que certains articles n'avaient pas de date complète de publication. Pour être insérée dans la base de données, une publication doit avoir une date comportant un mois constitué de trois lettres (jan,feb...) et d'une année composée de quatre chiffres. Dans le doute nous avons choisi d'attribuer le mois de janvier à ces références. Par contre, nous allions devoir marquer ces articles afin de prévenir les utilisateurs que nous ne connaissions pas le mois de publication de ces papiers. Nous nous sommes également aperçu qu'un auteur peut être orthographié de façons différentes. Par exemple : pour Jean-Philippe Berger nous pouvons trouver : "Berger,J." ou "Berger,J.P.". Un auteur peut aussi avoir plusieurs affiliations et une seule nationalité. Nous pouvions également imaginer que certains auteurs aient plusieurs nationalités. Et certains auteurs étaient sans affiliation. Il fallait donc que la base de données puisse s'adapter à ces cas particuliers. C'est Guillaume qui s'est occupé de résoudre ces problèmes.

3.2.3 Métadonnées

Une fois toutes les références insérées dans la base de données, il fallait reprendre chacune d'entre elles manuellement et attribuer les tags concernant les données complémentaires. C'est un travail que nous aurions voulu démarrer ensemble, Fabien et moi, mais le temps ne nous permettait pas de tout faire. Nous avons choisi de privilégié l'exploitation des données déjà inscrites dans la base. Attribuer les tags nécessite la lecture des abstracts, cela prend du temps et est un travail qui peut être délicat. Cela aurait été formateur pour moi du point de vue de l'astronomie mais l'objectif de ce stage était bel et bien l'informatique. C'est donc Peter Lawson et Fabien qui se sont partagés la tâche.

3.3 Accessibilité

3.3.1 Modification de olbin

La nouvelle base de données devait être facilement consultable par les astronomes. Il suffisait de modifier le code PHP de Olbin de façon à ce que la liste de publications écrite à la main soit remplacée par la liste des publications présentes dans la base de données. La partie du code correspondant à la liste des publications a donc été remplacée par le code présenté en annexe A. Bien-sûr, ce code que j'ai écrit avec avec l'aide de Fabien est là juste pour se faire une idée. Le but était surtout pour moi de comprendre le fonctionnement du HTML et du PHP. Dans la pratique il faudra réécrire cette page plus proprement pour pouvoir afficher les listes d'articles par année. Il faudra insérer dans le code une requête Sql permettant de sélectionner l'année. Il sera également intéressant de trier les listes par type de papier sur la page web, comme l'avait fait Peter Lawson sur l'ancienne page Olbin. Là encore une requête Sql permettra de trier les listes par type de papier. La page définitive sera discutée avec Peter Lawson.

A la fin de mon stage le prototype de la base de données se trouvait sur le site http ://jmmc.fr/bibdb/ ou son accès nécessitait une autorisation. La modification de Oblin allait bientôt faite afin de rendre la base de données consultable par tous.

3.3.2 L'ajout de références

Nous voulions que les chercheurs insèrent eux même leurs nouveaux papiers dans la nouvelle base de donnée dès leurs publications. Pour cela il fallait créer une page web indépendante permettant aux astronome d'insérer une références via son bibcode. Ainsi les données principales pourraient récupérées sur olbin à l'aide d'un script et les chercheurs mettraient eux même les tags. Sur cette page il pourraient aussi faire des commentaires et des suggestions. Il fallait aussi qu'ils puissent ajouter des tags auxquels nous n'aurions pas pensé ou correspondants à des instruments qui n'existaient pas lors de la création de la base de données par exemple. Il fallait être vigilant que l'on ne puisse pas ajouter un Tag qui existait déjà et qui n'aurait pas été vu ou qui aurait été orthographié différemment. La base de données allait donc être un outils évolutif pouvant sans cesse être améliorée et adapté aux besoins des chercheurs.

Le temps nous a manquer pour créer cette page durant ma période de stage. De plus, quelques détails restaient encore à régler concernant sa création.

4 Exploitation des données

Le premier prototype contenait simplement les listes d'articles de Olbin et leur données principales. Il n'y avait donc pas encore les métadonnées et pourtant nous pouvions déjà voir concrètement l'intérêt de créer cette base de données. Les nombreuses informations difficilement consultables sur l'ancienne page olbin étaient devenues facilement exploitables. Une simple requête Sql, nous permettait de savoir le nombre d'articles publiés depuis le début de l'interférométrie optique. Pour exemple le code en annexe B.1 nous a permit d'écrire une page PHP donnant le nombre

d'articles dans la base de données et de visualiser l'évolution du nombre de publications au cours des années. La page PHP correspondante est disponible sur le site du prototype (cité au paragraphe 3.3.1). L'accès à cette page nécessitant une autorisation, le résultat obtenu est présenté en annexe B.3 afin de donner une idée. Il était alors évident que pouvoir faire ce genre de statistiques est très intéressant d'un point de vue scientifique. Une fois tous les tags attribués nous allions pouvoir faire des statistiques sur les instruments les plus utilisés, les pays les plus actifs dans l'interférométrie optique, ou encore le type d'objet préférentiellement concernés par cette technique.

5 Bilan du projet

Mon stage s'est terminé il y a un peu plus de deux mois maintenant. Lorsque je suis partie, la base de données commençait à prendre forme. Elle est aujourd'hui accessible sur le serveur Jmmc à l'adresse suivante http ://apps.jmmc.fr/bibdb/olbin. Les problèmes qu'il reste à résoudre sont maintenant de l'ordre du détail. Un papier sera bientôt publié. Il sagit d'un poster qui permettra de faire connaître notre base de données. Les retours au sein du laboratoire sont très positifs et nous espérons que les remarques et les suggestions continuons afin d'améliorer encore la base de données.

6 Conclusion

Pour moi c'est un stage réussi. Je ne me suis jamais ennuyée et je venais avec plaisir au Laog malgré un semestre chargé. J'ai beaucoup appris et pas seulement de l'informatique. L'expérience et les conseils de Fabien m'ont permis d'avancer encore un peu plus dans la construction de mon projet professionnel. J'ai appris aussi que dans un projet comme celui-ci il faut savoir déléguer et travailler en équipe. C'est exactement se qu'a fait Fabien et le résultat est réussi. Il a eu l'idée de créer une base de données et il a su s'entourer des informaticiens du Laog qui lui ont permis de rendre ce projet concrétisable. Peter Lawson, à l'autre bout du monde, a travaillé avec Fabien sur la partie plus astronomique du problème et moi j'ai été là pour faire les petites choses pas trop difficiles mais pour lesquels il fallait que quelqu'un consacre un peu de temps. Je pense que l'objectif du magistère en nous faisant faire un stage en binôme était de comprendre se que c'est de travailler ensemble sur un même projet. Je n'ai pas fait ce stage en binôme néanmoins cet objectif est largement rempli. Lorsque j'ai su qu'un papier allait être publié au sujet de cette base de données et que j'allais faire partie de la liste des coauteurs, j'ai eu la confirmation que j'avais pu apporter ma pierre à l'édifice et c'était là mon objectif personnel. Même si se n'est pas moi qui ai fait le gros du travail sur cette base, j'ai pu aider à faire démarrer ce projet et j'en suis ravie. Je dis donc un grand merci à Fabien pour m'avoir proposé ce stage et m'avoir fait confiance. Merci aussi à Guillaume pour sa patience lorsqu'il a du m'expliquer plusieurs fois les même choses. J'espère pouvoir suivre l'évolution à moyen-long terme de la nouvelle base de données.

A Code de modification de Olbin

A.1 listPub.php

```
<?php
include_once('jmmc-web.inc.php');
jmmc_web_header("List of publications");
?>
<h1>List publications</h1>
<?php
// init $db
include_once('main.inc.php');

$articles = $db->GetActiveRecords("articles");
echo "<p>".count($articles)." articles in the database </p>";

$sql = "select * from articles";
$pager = new ADODB_Pager($db,$sql);
$pager->Render($rows_per_page=10);

jmmc_web_footer();
?>
```

A.2 main.inc.php

```
<?php
include_once('jmmc-web.inc.php');
include_once('/usr/share/php-adodb/adodb.inc.php');
include_once('/usr/share/php-adodb/adodb-pager.inc.php');
include_once('/usr/share/php-adodb/tohtml.inc.php');
require_once('/usr/share/php-adodb/adodb-active-record.inc.php');
session_start();
$db = NewADOConnection('mysql');
$db->Connect('localhost','bib','bib131','bib');
?>
```

B Exemple de statistiques possibles

B.1 stat.php

```
<?php
include_once('jmmc-web.inc.php');
jmmc_web_header("Interferometry Bibliography Database Statistics");
include_once('main.inc.php');

echo "<h2>Articles number</h2>";
```

```
$articles = $db->GetActiveRecords("articles");
echo "<p>".count($articles)." articles in the database </p>";

echo "<img src='dateplot.php'/>";

jmmc_web_footer();
?>
```

B.2 dateplot.php

```
<?php
require_once '/usr/share/phplot/phplot.php';
include_once('main.inc.php');

$data = array();

$sql = "SELECT DISTINCT (year) FROM articles ORDER BY year ASC";
$rs = $db->Execute($sql);
while (!$rs->EOF) {
    for ($i=0, $max=$rs->FieldCount(); $i < $max; $i++){
         $year = $rs->fields[$i];
$articles = $db->GetActiveRecords("articles", "year='$year'" );
$n = count($articles);
$data[]=array($year, $n);
}
    $rs->MoveNext();
}


$plot = new PHPlot(900, 500);
$plot->SetImageBorderType('plain');
$plot->SetPlotType('thinbarline');
$plot->SetDataType('text-data');
$plot->SetDataValues($data);
$plot->SetDataColors(array('blue'));
$plot->SetTitle('Articles number evolution');
$plot->SetYTitle('Number');
$plot->SetXTitle('Year');
$plot->SetLegend(array('All articles'));
$plot->SetYDataLabelPos('none');
$plot->SetXTickIncrement(5);
$plot->SetYTickIncrement(10);
$plot->SetPlotAreaWorld(0, 0, NULL,NULL );
$plot->SetXTickLabelPos('none');
$plot->SetXTickPos('none');
```

```
$plot->SetLineWidths(5);
$plot->DrawGraph();
?>
```

B.3 Résultat obtenu

```
Articles number
710 articles in the database
```