Ecole d'ingénieurs et d'architectes de Fribourg
Hochschule für Technik und Architektur Freiburg

Projet de diplôme

Infoscience
Spécification

Sylvain Egger

25 septembre 2007

INTRODUCTION

Infoscience & Invenio

Infoscience est une base de données de publications, rapports de recherche, thèses, travaux d'étudiants, cours, etc., provenant des facultés, laboratoires et chercheurs de l'Ecole Polytechnique Fédérale de Lausanne (EPFL).
En septembre 2007 Infoscience présente le profil de plus de 1.400 chercheurs et les publications de 180 laboratoires, sur les 240 qui composent l'EPFL. 60 000 publications y sont signalées, 12.000 Fulltexts y sont stockés, dont l'ensemble des thèses de l'EPFL. De plus Infoscience est destiné à devenir l'interface d'interrogation du catalogue collectif des bibliothèques de l'EPFL.
Cette base de données est liée à Invenio [1]. Invenio est une plateforme open source, développé et maintenue au CERN, qui sert de support au service Infoscience.

Infoscience fournit les contenus suivants

- Références et textes intégraux des publications scientifiques
- Description et identification des personnes et de leurs compétences
- Collections de documents numérisés ou numériques
- Le catalogue collectif des bibliothèques

Infoscience fournit les services suivants

- Interfaces de recherche Google-like et avancée
- Réutilisation des données par les chercheurs eux-mêmes
- Indexation des publications et profils de personne
- Interface de gestion de la qualité
- Conseil en droit d'auteur
- Intégration dans le moteur de recherche interne de l'EPFL

Curator

Afin d'assurer la qualité de données contenues dans Infoscience, un certain nombre de modules ont été développés afin de permettre de maintenir la qualité des données bibliographiques. Ces modules sont rassemblés sous le nom de « Curator » et sont disponibles à l'url suivante : http://infoscience.epfl.ch/quality/.

Actuellement ces modules sont en version Beta et ne sont pas implémentés dans une architecture adéquate.
Voici une liste des modules qui devrait être disponible dans la version finale du Curator.

- Uploader un fichier de notices
- Gestion des doublons
- Archivage d'un laboratoire
- Définir le format d'autorité d'un nom d'auteur
- Transféré les publications d'une personne
- Modifier le type d'une notice
- Supprimer une notice d'un laboratoire
- Ajouter ou modifier un journal scientifique à la base de référence
- Ajouter un livre à la bibliothèque d'un laboratoire

CAHIER DES CHARGES COMPLET DE CURATOR

Il existe déjà un cahier des charges [2] décrivant l'entier des modules du Curator.
Ce cahier des charges peut se résumer sous forme de Use Case.

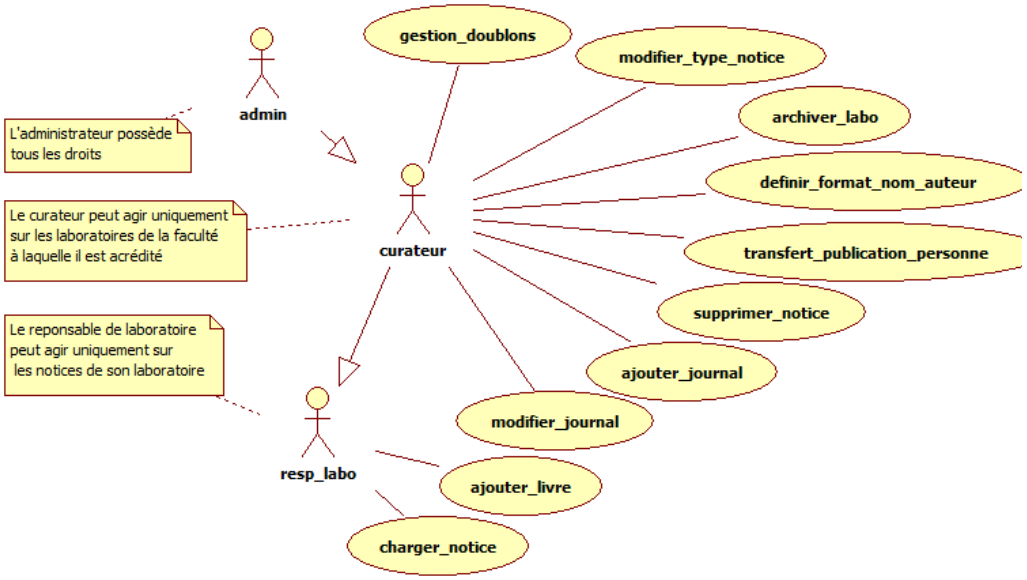


Figure 1: Diagramme Use Case des modules Curator

Ces différents modules agissent directement sur des notices présentes dans la base de données d'Invenio. Mais cela pourrait être une autre plateforme bibliographique.

Tâches

Introduction

Le travail de ce projet consiste à préparer un socle sur lequel ces modules viendront se greffer. Ensuite, une fois l'architecture de base créée, les différents modules pourront être réalisés. Cependant, durant ce projet, il n'est pas question de réaliser tous ces modules mais uniquement un ou deux modules si le temps le permet.

Définir une architecture

Dans un premier temps, il faudra réfléchir à une architecture solide. Ce socle de base devra pouvoir accueillir les différents modules du Curator d'une manière la plus générique possible. C'est à dire que cette architecture devra également prendre en compte le fait que le Curator n'est pas exclusivement destiné à Invenio. Au contraire, le Curator doit être un module qui pourrait s'adjoindre à d'autres plateformes semblables à Invenio.

Cette architecture devra gérer la communication entre Invenio et le Curator afin de permettre aux modules de pouvoir traiter des notices de la meilleure des façons selon leurs besoins. Autrement dit, elle devra fournir aux modules des méthodes pour communiquer avec Invenio.

Exemple possible du flux de communication

Liste des tâches à réaliser pour l'architecture

1. Définir si la plateforme Curator devra ou non être placée sur la même machine que la plateforme bibliographique *(Invenio dans notre cas)*

2. Définir la façon dont les flux de données vont transiter entre Invenio et Curator. Définir les librairies, les méthodes qui vont être utilisées pour former les flux de données. Par exemple, l'utilisation du XML MARC et la forme des notices XML MARC.

3. Définir exactement les fonctions du Curator qui seront fournies aux modules qui viendront se greffer au-dessus.

Développer un module sur la nouvelle architecture

Dans un second temps, une fois l'architecture définie et implémentée, il faudra réaliser un des modules du Curator afin de tester et valider l'architecture.

Le module à développer sera choisi en fonction du temps à disposition.

Cependant, le module « uploader un fichier de notice » qui permet de fournir un fichier au format XML MARC pour insérer ou mettre à jour des notices devrait forcément être implémenté afin de démontrer la communication entre Invenio et Curator.

En effet, le chargement de notices est une étape présente dans presque tous les modules. Pas toujours exactement avec les mêmes informations dans la notice mais l'idée du transfert d'un flux XML MARC reste la même.

Liste des tâches à réaliser pour la création d'un module

1. Respecter le prototype design du module [4]
2. Respecter le processus défini dans le cahier des charges de Curator
3. Optimiser la gestion de la communication avec Invenio

Organisation

Administratif

Divers documents doivent être rédigés durant le projet. En voici la liste exhaustive.

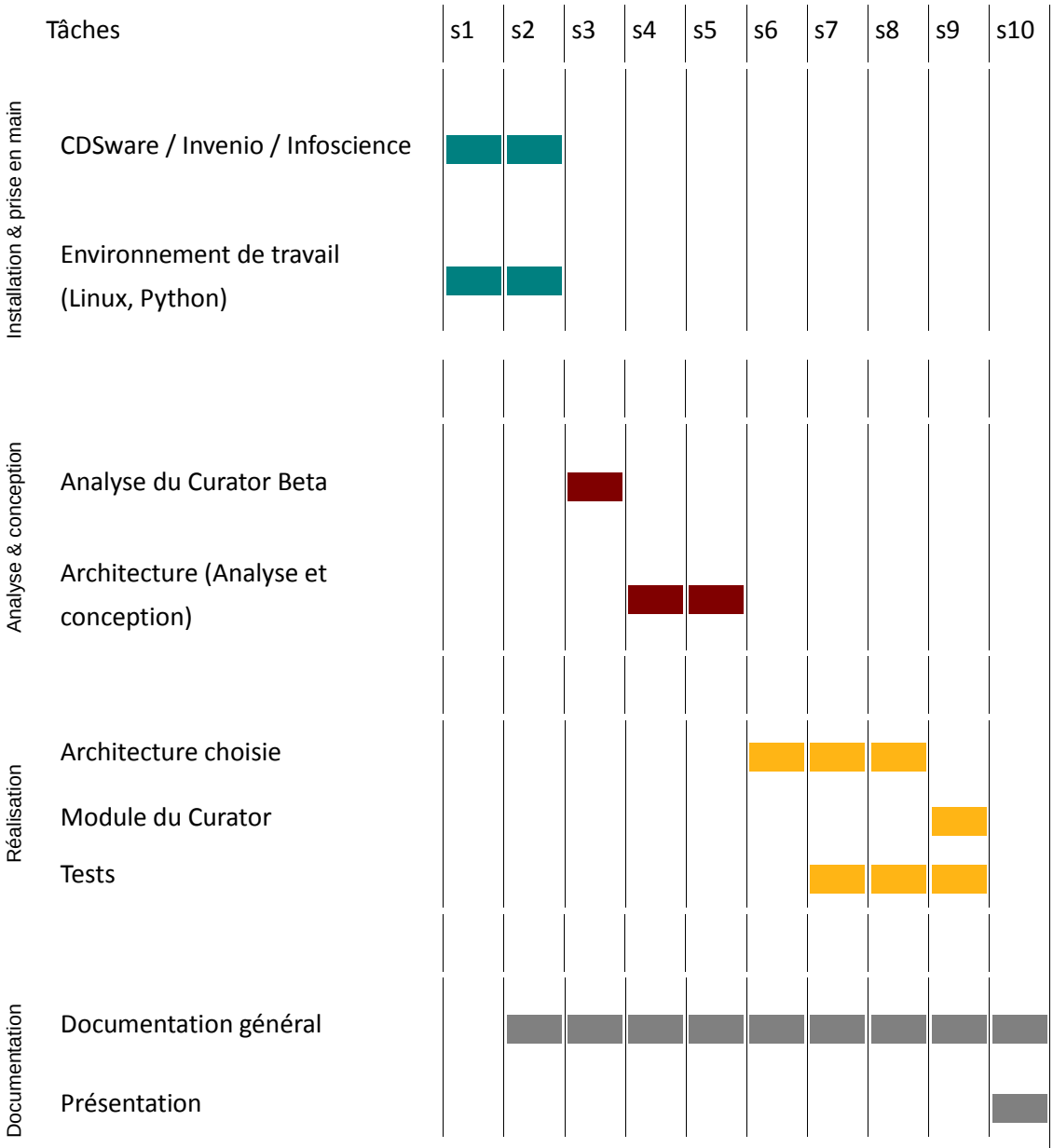
- Un rapport technique décrivant toutes les étapes du projet
- Des Pvs pour chaque séance tenue

Il a également été défini que tous les lundis, aura lieu une réunion afin d'observer l'avancement du projet et de définir les points à traiter pour la semaine.

Puis une fois toutes les deux semaines au minimum et plus si nécessaire, une réunion sera agenda à Fribourg avec les responsables. La convocation est à effectué par email avec l'ordre du jour.

L'étudiant se charge également de contacter ses experts afin de les tenir au courant de l'état du projet et de les rencontrer si nécessaire.

Planification



Références

[1] : Invenio est un système de bibliothèque numérique, une suite d'applications qui fournit une plateforme et des outils pour créer et gérer un serveur de publications numérique.
Plus d'informations : *http://cdsware.cern.ch/invenio/*

[2] : Pierre Crevoisier et Gregory Favre ont défini un cahier des charges de Curator. Celui-ci est en version 1.0 depuis Août 2007.
Plus d'informations : *http://empc51.epfl.ch/infoscience/Curator*

[3] : XML MARC désigne un format de données permettant d'informatiser les catalogues de bibliothèques dans un format XML défini. Il se présente à l'écran comme une succession de champs de données, appelée grille MARC, de longueur variable portant chacun une étiquette (un nombre de 3 chiffres)
Plus d'informations : *http://www.loc.gov/marc/marcxml.html*

[4] : Un prototype design décrivant les interfaces web pour les modules de Curator a déjà été défini.
Plus d'informations : *http://empc51.epfl.ch/infoscience/Curator*

Glossaire

doublon	Un doublon est une paire (ou plus) de notices, ayant le même contenu, étant présentes deux fois (ou plus) dans Infoscience.
laboratoire	Un laboratoire est un groupe de recherche qui envoi des publications sur Infoscience. Deux laboratoire peuvent travailler ensemble sur une publication et la publier chacun de leur coté, ce qui engendrera un doublon.
nom d'auteur	Un nom d'auteur est le champ auteur dans une notice. Pour une même personne il peut-être différent. Par exemple : Egger, S ou Egger, Sylvain. Le but est de regrouper le nom d'auteur synonyme pour améliorer la recherche par auteur.
notice	Une notice est une publication Infoscience. Elle contient toutes les informations telles que le titre, les auteurs, le laboratoire, les fichiers joints,...etc.