Project proposal Bachelor Degree Project 20171215

Project Title: Connecting Silos

Authors

Shiva Besharat Pour Emmylundsvägen 3/1605, SOLNA shivabp@kth.se 0046-760 65 61 53	Qi Li Björksundsslingen 23 5tr 12431 Bandhagen Stockholm qi5@kth.se 0046-707160955

The students will divide the responsibilities with respect to the different interfaces that need to be defined and understood.

Organization and Supervisor

The project will be organized as a two person project building upon previous work that has already been done by Gerald Q. Maguire Jr. regarding the custom files in Canvas for thesis projects and the existing work he has already done in parsing PDF documents.

Organization: KTH, EECS, CoS, RSlab

Supervisor: Ander Västberg

Examiner's name: Gerald Q. Maguire Jr.

Keywords

RESTful APIs, Canvas, DiVA, Calendars, data mining

Background

The students need to be able to understand RESTful APIs, network services, and basic concepts of data mining and web page scraping/fill in.

Problem statement

With the increasing digitization of the academic processes, there is a need to integrate the various systems that have heretofore been implemented as separate silos.

Problem

Extending Canvas to automate pushing a submitted thesis that has been accepted by the examiner to DiVA (complete with transfer of meta data, production of covers, uploading of the thesis, and marking whether it is to be immediately public of not), similarly pushing the relevant information from a

beta draft and the date, time, place of the oral presentation to a Calendar system to automate the announcement of the event.

Purpose

When a student has submitted a beta-draft and the examiner has scheduled the oral presentation it should be possible for the examiner to push a button in Canvas and have the calendar entry made automatically using information in the draft and Canvas. Similarly when the examiner has approved the thesis, the examiner should be able to push a button in Canvas and have the front and back covers added to the thesis and the meta data and thesis put into DiVA. If the student has given approval for the full text to be made publicly available this should be reflected in the DiVA s submission.

Goal(s)

The existing manual process of putting a thesis into DiVA takes ~1 hour of time. In 2016, there were 2373 student theses entered into DiVA[1]. If we assume 40 weeks per year at 40 hours per week, this represents ~1.5 people (over all of KTH)! Automating this process would both save time and increase the accuracy of the data, as well as making the thesis available to the public sooner

Tasks

The following tasks must be addressed in the course of this thesis project:

1. Learn how to get information out of and put information into Canvas [1] [This means understanding how a RESTful API works.].
2. Learn how to parse a thesis document in PDF (or perhaps even Word format) to extract key meta information: title, subtitle, author(s), school, degree, date, abstract and keywords in both English and Swedish (and possibly others), and number of preface pages and number of pages in the thesis [This means understanding how to do data minim from a PDF document using a tool such as pdfssa4met [2].].
3. Learn how to generate a cover using the KTH cover generator [3] using a POST with the relevant values. Then combine the front and back covers with the approved thesis (for example using PyPDF2 [4]).

1 Data extracted by creating a feed from DiVA for the year 2016 for student theses.

4. Learn how to insert this data into Digitala vetenskapliga arkive (DiVA) [5–7] [This means looking either a headless browser approach, such as selenium or GreaseMonkey [8], directly entering the data or importing a Metadata Object Description Schema (MODS) [9] file, alternatively defining together with DiVA an API.]
5. Learn how to insert this data into Calendar system [This means looking either an API such as time edit or defining an API to Polopoly.]
6. Check data for correctness and consistency [An eternal problem in digitization is ensuring that data is consistent.]
7. To facilitate data consistency, information about the examiner and internal academic adviser(s) should be collected automatically (including information such as their ORCID ID [10], KTH ID, etc.)

Method

An empirical method will be used to evaluate a series of prototypes in order to incrementally develop and evaluate the proposed services.

Milestone chart (time schedule)

< Milestone chart illustrates the project timeline and when will particularly meaningful points, referred to as milestones, are to be completed.
Also mention the deliverable for each of these milestones. >

Risks, Consequences and Ethics

There exists a Canvas API to be able to get the thesis or draft thesis. It is not yet known how to identify the student for whom the schedule “button” or report to DiVA button should apply. It is not yet known how to put the data into DiVA, but there is the ability to import a MODS record for each thesis (just as for any other publication), but it is not known how to insert the full-text of the thesis or indicate that it can be public or not. Similarly, there is an API to TimeEdit and to Canvas for announcing events in a calendar, but it is not clear if there is a means to do this via the Polopoly calendar for EECS (however, there is evidence in the form of a web interface to do such announcements at the Karolinska Insitute). Note that all theses in Sweden are public, so there is not any problem about confidential material with regard to

the meta data or thesis itself once the examiner has approved it. The only potentially confidential material is the name(s) of industrial advisers, who may or may not want their name and affiliation to be public with respect to the thesis project.

Summary

The main result should be a prototype of a tool that an examiner can use to push data from Canvas into a calendar system to announce a thesis presentation and when the final thesis is done to push the relevant data and thesis to DiVA.

Reference(s)

[1] Instructure Inc., ‘Canvas LMS REST API Documentation’, 14-Dec-2017. [Online]. Available: https://canvas.instructure.com/doc/api/index.html. [Accessed: 15-Dec-2017]

[2] Elias Kunnas, PDF Structure and Syntactic Analysis for Metadata Extraction and Tagging: https://code.google.com/p/pdfssa4met/. 2017 [Online]. Available: https://github.com/eliask/pdfssa4met. [Accessed: 15-Dec-2017]

[3] Niklas Olsson, ‘KTH Book Cover Generator’, *Skapa omslag till examensarbete*. [Online]. Available: https://intra.kth.se/kth-cover. [Accessed: 15-Dec-2017]

[4] Tim Arnold, ‘Manipulating PDFs with Python | Python’, *Binpress*, 06-Nov-2014. [Online]. Available: http://www.binpress.com/tutorial/manipulating-pdfs-with-python/167. [Accessed: 15-Dec-2017]

[5] Enheten för digital publicering (EPC) Uppsala University Library, ‘DiVA portal is a finding tool for research publications and student theses written at the following 47 universities and research institutions.’ [Online]. Available: http://www.diva-portal.org/smash/aboutdiva.jsf. [Accessed: 15-Dec-2017]

[6] Enheten för digital publicering (EPC), Uppsala University, ‘EPC Homepage’. [Online]. Available: http://epc.ub.uu.se/. [Accessed: 15-Dec-2017]

[7] Karin Meyer Lundén,Enheten för digital publicering (EPC), Uppsala University, ‘DiVA – kort systembeskriving’, 24-Apr-2013. [Online]. Available: https://wiki.epc.ub.uu.se/download/attachments/2064421/DiVAintro.pdf. [Accessed: 15-Dec-2017]

[8] Anthony Lieuallen, ‘Greasespot’, 11-Dec-2017. [Online]. Available: https://www.greasespot.net/. [Accessed: 15-Dec-2017]

[9] U.S. Library of Congress, ‘Metadata Object Description Schema: MODS (Library of Congress Standards)’. [Online]. Available: http://www.loc.gov/standards/mods/. [Accessed: 15-Dec-2017]

[10] ORCID, Inc., ‘ORCID’. [Online]. Available: https://orcid.org/. [Accessed: 15-Dec-2017]