DEGREE PROJECT XXX, FIRST CYCLE
STOCKHOLM, SWEDEN 2018

Connecting Silos

Automation System for Thesis Processing in Canvas and Diva

QI LI;Shiva BESHARAT POUR

Kommenterad [GQMJr1]: The front and back covers should be made by the examiner when the thesis is ready to be approved.

This title is take from the "Title" properties of the document.

Kommenterad [GQMJr2]: The subtitle is taken from the "Subject" properties of the document. (Why Microsoft calls it this is not known.)

Kommenterad [GQMJr3]: This field is taken from the "Author" properties of the document.

KTH ROYAL INSTITUTE OF TECHNOLOGY
Electrical Engineering and Computer Science

Connecting Silos

Automation System for Thesis Processing in Canvas and Diva

Qi Li;Shiva Besharat Pour

2018-04-03

Bachelor'sThesis

Examiner
Gerald Q. Maguire Jr.

Academic adviser
Anders Västberg

KTH Royal Institute of Technology
School of Electrical Engineering and Computer Science (EECS))
Department of Communication Systems
SE-100 44 Stockholm, Sweden

Kommenterad [GQMJr4]: This is the inside title page – **the student should write this page and the abstract** as the first thing that they do after deciding to do a thesis project.

This title is take from the "Title" properties of the document.

Abstract

Write an abstract with the following components:

- What is the topic area? (optional) Introduces the subject area for the project.
- Short problem statement
- Why was this problem worth a Master's thesis project? (i.e., why is the problem both significant and of a suitable degree of difficulty for a Master's thesis project? Why has no one else solved it yet?)
- How did you solve the problem? What was your method/insight?
- **Results/Conclusions/Consequences/Impact**: What are your key results/conclusions? What will others do based upon your results? What can be done now that you have finished - that could not be done before your thesis project was completed?

As the era of digitalization dawns, the need to integrate separate silos into a synchronized connected system is becoming of ever greater significance. This thesis focuses on the Canvas Learning Management System (LMS) and the Digitala vetenskapliga arkive (DiVA) as examples of separate silos.

The thesis presents several methods of automating document handling associated with a degree project. It exploits the fact that students will submit their thesis to their examiner via Canvas. Canvas is the LMS platform used by students to submit all their coursework. When the examiner approves the thesis it will be archived in DiVA and optionally published via DiVA. DiVA is an institutional repository used for research publications and student theses.

When manually archiving and publishing student theses via DiVA several fields need to be filled in. These fields is provide meta data for the thesis itself. The content of these fields (author, title, keywords, abstract, …) can be used when searching via the DiVA portal. While it might not seem like a massive task to enter this meta data for an individual thesis; however, given the number of theses that are submitted every year, this process takes a significant amount of time and effort. Moreover, it is important to enter this data correctly – which is difficult when manually doing this task. Therefore, this thesis project seeks to automate this process for future theses.

The solution that presented in this thesis will parse PDF documents and use other information from the LMS in order to automatically generate a cover for the thesis and fill in the required DiVA meta data. This data will also be inserted into a calendar system to provide an announcement for the student's thesis presentation. Moreover, the data will be checked for correctness and consistency.

Manually filling in DiVA fields in order to publish theses has been a quite demanding and time-consuming process. Thus, there is often a delay before a thesis is published via DiVA. Therefore, this thesis project's goal is to provide KTH with an

Kommenterad [GQMJr5]: Keep in mind that most of your potential readers are only going to read your title and abstract. This is why it is important that the abstract give them enough information that they can decide is this document relevant to them or not. Otherwise the likely default choice is to ignore the rest of your document.

A abstract should ***stand on its own***, i.e., no citations, cross references to the body of the document, acronyms must be spelled out, …

Write this early and revise as necessary. This will help keep you focused on what you are trying to do.

Kommenterad [gqmjr6]: Use about ½ A4-page - 250 and 350 words.

Kommenterad [gqmjr7]: The presentation of the results should be the main part of the abstract.

Kommenterad [gqmjr8]: Note that one you have spelled it out and introduced the abbreviation you simply use the abbreviation for the rest of the abstract.

Kommenterad [gqmjr9]: Note that the DiVAportal is the search interface, while the actual database is DiVA. See for example https://sv.wikipedia.org/wiki/Diva_(digitalt_arkiv)

automated means to handle thesis archiving and publication via DiVA, while doing so faster, more efficiently, and with fewer errors.

Kommenterad [gqmjr10]: These can be metrics that you can use to assess your solution.

Keywords

5-6 keywords

Kommenterad [gqmjr11]: Choosing good keywords can help others to locate your paper, thesis, dissertation, … and related work. Choose the most specific keyword from those used in your domain, see for example:
ACM's Computing Classification System (2012)
(2014) IEEE Taxonomy
Mechanics:
- •The first letter of a keyword should be set with a capital letter and proper names should be capitalized as usual.
- •Spell out acronyms and abbreviations.
- •Avoid "stop words" - as they generally carry little or no information.
- •List your keywords separated by commas (",").

Since you should have both English and Swedish keywords - you might think of ordering them in corresponding order (i.e., so that the n^{th} word in each list correspond) - thus it would be easier to mechanically find matching keywords.

Sammanfattning

Eftersom den digitaliserade eran förbättras, blir behovet av att integrera separata silor i ett synkroniserat anslutet system allt viktigare. Denna thesis kommer att fokusera på Canvas Learning Management System och Digitala vetenskapliga arkive (DiVA) Portal som de separata silorna.

Under hela denna thesisrapporten presenteras metoder för att automatisera Canvas beträffande publicering av inlämnad thesis till. Canvas är för närvarande den plattform som används av Kungliga Tekniska högskolan (KTH) där studenter lämnar in allt sitt kursarbete från små volymer till stora volymer som t.ex. thesisrapporter. DiVA är en institutionell förvaringsplats som används för

thesisrapporter. Därför måste flera fält fyllas i när man publicerar thesisrapporter från Canvas till DiVA manuellt.

Att fylla i DiVA-fält för att publicera thesisrapporter kanske inte verka som en stor uppgift om det bara fanns några få antal thesisrapporter. Men med tanke på det stora antalet thesisrapporter som varje år skickas till Canvas, kommer denna process att ta en betydande tid och mankraft som inte ska underskattas. Därför kommer detta bachelorsprojekt att fokusera på att ta itu med detta problem för framtida dessa inlägg.

Lösningen som presenteras i hela denna rapport kommer att bestå av metoder för att analysera Portable Document Format (PDF) -dokument och skapa ett omslag med hjälp av den analyserade informationen och fylla i de önskade DiVA-fälten automatiskt. Dessa data kommer också att införas i ett kalendersystem och kommer att kontrolleras både automatiskt och manuellt, för korrekthet och konsistens.

Manuell fyllning av DiVA-fält för att publicera thesisrapporter har varit en ganska krävande och tidsbesparande process för KTH under de senaste åren, där ingen automatisering för denna process tillhandahölls. Detta har lett till att en uppsättning av thesisrapporter som väntar på att publiceras på DiVA och inte tillräckligt med tid för att någon anställd tar ansvar för att ta hand om problemet. Därför kommer det här projektet att kunna utrusta KTH:s personal med ett automatiskt publiceringssystem, vilket ger effektivitet till thesisrapporternas publikation i DiVA.

Kommenterad [GQMJr12]: All theses at KTH are required to have an abstract in both English and Swedish.

If you are writing your thesis in English, you can leave this until the final version. If you are writing your thesis in Swedish then this should be done first – and you should revise as necessary along the way.

If you are writing your thesis in English, then this section can be a summary targeted at a more general reader. However, if you are writing your thesis in Swedish, then the reverse is true – your abstract should be for your target audience, while an English summary can be written targeted at a more general audience.
⇒ This means that the English abstract and Swedish sammnfattning
or Swedish abstract and English summary need not be literal translations of each other.

The abstract in the language used for the thesis should be the first abstract, while the Summary/Sammanfattning in the other language can follow.

Exchange students many want to include one or more abstracts in the language(s) used in their home institutions to avoid the neeed to write another thesis when returing to their home institution.

Nyckelord

5-6 nyckelord

Kommenterad [gqmjr13]: nyckelord som beskriver innehållet i uppsatsrapporten

Acknowledgments

I would like to thank xxxx for having yyyy.

Stockholm, Month Year
Qi Li;Shiva Besharat Pour

Kommenterad [GQMJr14]: It is nice to acknowledge the people that have helped you. It is also necessary to acknowledge any special permissions that you have gotten – for example getting permission from the copyright owner to reproduce a figure. In this case you should acknowledge them and this permission here and in the figure's caption.

[zNote: If you do not have the copyright owner's permission, then you cannot use any copyrighted figures/tables/… .]

Table of contents

List of Figures

List of Tables

List of acronyms and abbreviations

API	Application Programming Interface
CMS	Content Management System
DiVA	Digitala vetenskapliga arkive (Swedish)
EECS	School of Electrical Engineering and Computer Science
ISRN	International Standard Technical Report Number
ISSN	International Standard Serial Number
IT	Information Technology
KTH	KTH Royal Institute of Technology (English) / Kungliga Tekniska högskolan (Swedish)
LMS	Learning Management System
MODS	Metadata Object Description Schema
ORCID	Open Researcher and Contributor ID
PDF	Portable Document Format
XML	Extensible Markup Language

Kommenterad [GQMJr15]: The list of acronyms and abbreviations should be kept in alphabetical order based on the spelling of the acronym or abbreviation.

Note that this is formatted as a table, so when working on the document you can turn on the borders to make the table easier to work with. When submitting the final version of the thesis turn of the borders.

1 Introduction

This chapter describes the specific problem that this thesis addresses, the context of the problem, the goals of this thesis project, and outlines the structure of the thesis.

Kommenterad [gqmjr16]: The first paragraph after a heading is not indented, all of the subsequent paragraphs have their first line indented.

In order to achieve efficiency, it is desirable to automate routine tasks that are demanding with respect to time and effort when done manually. This thesis presents the design, implementation, and evaluation of an automation solution applied to processing some elements of degree projects. The aim is to provide increased efficiency, increased accuracy, and reduce the time and effort needed by *all* involved. Note that the goal is with respect to everyone who is involved with the process, thus it includes the students who author the thesis, the faculty (as an examiner), and administrative staff.

Give a general introduction to the area. (Remember to use appropriate references in this and all other sections.)

1.1 Background

Present the background for the area. Set the context for your project – so that your reader can understand both your project and this thesis. (Give detailed background information in Chapter 2 - together with related work.)

Sometimes it is useful to insert a system diagram here so that the reader knows what are the different elements and their relationship to each other. This also introduces the names/terms/... that you are going to use throughout your thesis (be consistent). This figure will also help you later delimit what you are going to do and what others have done or will do.

Canvas is a learning management system (LMS) used by many schools and institutions, for assignments and coursework. One of the main purposes of Canvas is for the teachers to create coursework/assignments, and for students to submit their work and receive a grade. It is quite well organized and automated when it comes to handling student submissions. It is also useful when it comes to its automation regarding organizing the assignments consistently concerning their priority. DiVA is a publishing system for research and student theses. More detailed information about the Canvas LMS and DiVA portal will be provided in Sections 2.3 and 2.4 (respectively).

KTH Royal Institute of Technology (here after simply KTH) uses DiVA as an archive for student theses and publishes approved student theses that have been submitted via Canvas to DiVA. However, to doing so, specific meta data has to be entered into DiVA. Currently, this meta data is entered via fields presented using a web interface to the DiVA Portal. These fields need to be filled in with information about the thesis (title, abstracts, keywords, number of pages, etc.), the student, the examiner, advisers, the defense, etc. The thesis is uploaded to DiVA for archiving and

optionally the full text of the thesis is published via DiVA. This entire process is currently being done manually. This requires a significant amount of time (roughly an hour per thesis) and effort for staff members to enter the meta data and thesis into DiVA. Moreover, before a thesis can be uploaded to DiVA it has to be assigned a report number, the cover made, and attached to the thesis.

Currently, automation is lacking when it comes to connecting Canvas to other digital platforms, such as Digitala vetenskapliga arkive (DiVA) .

Kommenterad [gqmjr17]: Note that one you have spelled it out and introduced the abbreviation you simply use the abbreviation for the rest of the thesis.

1.2 Problem

Longer problem statement

If possible, end this section with a question as a problem statement.

If the number of theses that are submitted were few, the time that the process of entering the theses into DiVA takes would be insignificant. However, considering the number of these that are submitted is high (see Table 1-1 and Table 1-2), thus some problems arise regarding the efficiency of the complete process. It is worth mentioning that, especially towards the end of every academic year, many students are submitting their thesis or dissertation. Therefore, not only are there a large number of theses to enter into DiVA, but the work is very concentrate in a small part of the year.

Table 1-1: Number of degree project reports in DiVA for all of KTH

Year	Total number	Full-text in DiVA	Full-text not available in DiVA
2017	2287	2053	234
2016	2376	2182	194
2015	2601	2316	285
2014	2384	2050	334
2013	2356	2035	321
2012	2500	1873	627
2011	2282	1640	642
2010	504	486	38

Table 1-2: In 2017, School of Electrical Engineering and Computer Science (EECS) had 697 theses (24 without full text)

Organisation	Number
School of Computer Science and Communication (CSC)	338
School of Information and Communication Technology (ICT)	154
School of Electrical Engineering (EES)	47
Electric Power and Energy Systems	30
Automatic Control	29
Media Technology and Interaction Design, MID	18
Information Science and Engineering	17
Electromagnetic Engineering	14
Space and Plasma Physics	10
Robotics, perception and learning, RPL	10

Manually extracting the meta information required by DiVA for each thesis is quite repetitive, demanding, and takes an unnecessarily large amount of time. This work can take months, whereas automating it would only require a press of a button by the examiner when approving the thesis and take only a matter of seconds for a computer to complete the process. Therefore, the main problem that this thesis project will try to solve is "How can approved student theses submitted via Canvas be automatically entered into DiVA?".

1.3 Purpose

State the purpose of your **thesis** and the purpose of your **degree project**.

Describe who benefits and how they benefit if you achieve your goals. Include *anticipated* ethical, sustainability, social issues, etc. related to your project. (Return to these in your reflections in Section 3.4.)

Kommenterad [gqmjr18]: This section should not begin "The project is about" even though this can be included in the purpose section. If so, state the purpose of the project *after* purpose of the thesis).

The purpose of this bachelor degree project is to design, implement, and evaluate a system to automate the entry of an approved thesis into DiVA. The project that will also try to automate the event-creation to announce student's oral presentation as the requirements are quite similar regarding parsing data from a submitted document and filling in fields to create an event in the university's Calendar system. Thus, the repetitive task of extracting information from theses submitted via Canvas and using this information enter the theses into DiVA as well as creating Calendar

events will be done automatically as soon as the appropriate button is pushed (by the responsible person).

The solutions to use for this automation will be thoroughly presented and described in sufficient detail for others to utilize. The methods used for solving the problem will be evaluated and compared to existing methods (if any). The algorithm developed for the automation will be presented. A number of tests will be carried out to demonstrate the correctness and consistency of the result of the algorithm. Moreover, an estimate of the time saved with the introduction of this automation will be made. This saved time can be used for other tasks that actually require human interaction (for example, better supporting the advising of students).

1.4 Goals

State the goal/goals of this degree project.

In addition to presenting the goal(s), you might also state what the deliverables and results of the project are.

Kommenterad [gqmjr19]: Note that in the literature study and even the alpha draft, these are your expected goals, deliverables, and results – which may change over the course of the project – hence you will revise this in the final report to describe what you actually achieved, delivered, and produced as results.

The goal of this degree project is to automate the processing of taking a thesis submitted via Canvas and entering it into DiVA or as a Calendar event in the case of oral presentations. This has been divided into the following two sub-goals:

1. Once an examiner has schedule an oral presentation, the extension to Canvas will automatically extract the relevant information needed to create a Calendar event for a given degree project presentation based upon the submitted beta draft and the time and place of the presentation.
2. Once an examiner has approved a thesis submitted via Canvas the relevant information will be extracted from the thesis itself and combined with other data that is available in Canvas to automate the full process of publishing theses via DiVA.

Achieving the above subgoals should provide greater efficiency that the current manual process for theses publication and oral presentation event creation.

1.5 Research Methodology

Introduce your choice of methodology/methodologies and method/methods – and the reason why you chose them. Contrast them with and explain why you did not choose other methodologies or methods. (The details of the actual methodology and method you have chosen will be given in Chapter **Fel! Hittar inte referenskälla.**. Note that in Chapter 3, the focus could be research strategies, data collection, data analysis, and quality assurance.)

In this section you should present your philosophical assumption(s), research method(s), and research approach(es).

The research method that this thesis will use is qualitative research. Qualitative research means the research is primarily exploratory research [1]. The qualitative research that is carried out for this thesis project will focus on understanding the

reason, opinions, and motivation [1] for the structure of the data inside a Portable Document Format (PDF) file and the methods to insert and extract data, to and from the Canvas LMS. In particular, we need to know how to parse a PDF document in order to extract the relevant data (such as title, abstracts, and keywords). This action will be followed by generation of a cover for the thesis, as well as combining the front and back covers with the thesis. Eventually, research will be carried out to check that the data automatically entered into DiVA based upon the extracted information is correctness and consistent. This correctness and consistency will be compared to previously data manually entered into DiVA.

This thesis project is generally based upon parsing information from documents and inserting the extracted data into the relevant fields of records in other systems, hence it is about connecting what are today separate silos. The details of this parsing and extracting will be described later in the thesis.

An implementation choice is which programming language will be used and what algorithm is best to extract the data from the relevant source. In this context, best can be evaluated in terms of efficiency – but it remains to be defined as to whether this is with respect to development efficiency or run-time efficiency (as we do not yet know how much time it will take to do the desired parsing and processing of the data).

The code provided by previous work (described in Chapter 2) is written in python, hence it would be simpler to implement the algorithm for this project if it too were written in python. Python is an interpreted high-level programming language for general-purpose programming [2]. How the algorithm is implemented will also depend on the Canvas Application Programming Interface (API) and how we will interact with DiVA.

1.6 Delimitations

Describe the boundary/limits of your thesis project and what you are explicitly **not** going to do. This will help you bound your efforts – as you have clearly defined what is **out of the scope** of this thesis project. Explain the delimitations. These are all the things that could affect the study if they were examined and included in the degree project.

1.7 Structure of the thesis

Kommenterad [gqmjr20]: Exclude the first chapter , references, and appendix/appendices.

Chapter 2 presents relevant background information about xxx. Chapter 3 presents the methodology and method used to solve the problem. …

2 Background

Kommenterad [GQMJr21]: When you do your literature study, you should have a nearly complete Chapters 1 & 2.

You may also find it convenient to introduce the future work section into your report early – so that you can put things that you think about but decide not to do now into this section.

Note that later you can move things between this future work section and what you have done as you may change your mind about what to do now versus what to put off to future work.

The Connecting Silo project is meant to simplify and automate the degree project administrative processes. This chapter begins with a rough description of the workflow in a 1st cycle degree project to clarify the overall administrative process for a degree project. As the project will be tested and developed based on the current version of the Canvas LMS running at KTH, DiVA, and the Polopoly platform (used for Calendar events), these three systems are introduced in Sections 2.2-2.4. This will be followed by introductions to details of these three systems, how one can interact with them, and background information about some of the tools to be used. The chapter ends with section on related work (Section 2.18), analysis of reliability and validity (Sections 2.19 and 2.20), and concludes with a discussion of the material presented in this chapter (Section 2.21).

What does a reader (another x student -- where x is your study line) need to know to understand your report?

What have others already done? (This is the "related work".) Explain what and how prior work / prior research will be applied on or used in the degree project /work (described in this thesis). Explain why and what is not used in the degree project and give valid reasons for rejecting the work/research.

2.1 Workflow of 1st cycle degree project at KTH

A Bachelor's thesis (is the result of a first cycle degree project). This degree project is done as a course worth 15 credits. The project is typically done at the end of the last study year of students in the first cycle [3, 4].

As Figure 2-1 demonstrates, in order to be able to register for the degree project course a student need to meet certain university requirements. Once they have met these prerequisites, then depending on the cycle and program of study of an individual student, the student will register for the thesis course under different course codes. The student will be added to a Canvas course for degree projects of a given cycle (in this work we will only consider 1st and 2nd cycle degree projects). Next the student will depending on their education program be placed by an administrator (or potentially automatically) into a section within this course. During the first cycle, the student will proceed with their thesis project either in a group of two people or alone [3, 4]. Note that 2nd cycle students do their degree project individually.

Once the student has been added to the Canvas course, then each student or group of students is required to submit a project proposal via Canvas. This project proposal describes the proposed project and methodology that will be used in the project. The student may also fill out a survey to provide additional information about the proposed degree project (such as suggested examiners, supervisors, whether they give their approval for the full text of the thesis to be published via DiVA, etc.). If the proposed project meets the requirements of a given program, then the Program Administrator will assign an examiner and supervisor for the student

based on the topic of the proposal. The potential examiner will be recorded in the Canvas GradeBook (for details see Section 2.6). This examiner will read the project proposal and decide if the proposal has sufficient quality to be accepted and if they are a suitable examiner. If the proposal has low quality, then the examiner will ask the student to improve the proposal based on the examiner's feedback. This process continues until the proposal is accepted by the examiner. Additionally, the assigned examiner may reject being the examiner for a particular student (typically because the topic is out of the scope of this examiner's expertise, the examiner is already overloaded, there is a conflict of interest, etc.). In this case, the Program Administrator will assign another examiner to this student [3, 4]. There are two potentially automated parts of this workflow and they have been tagged in Figure 2-1. The Connecting Silo project will support the automation for these 'potentially automated' parts of the workflow.

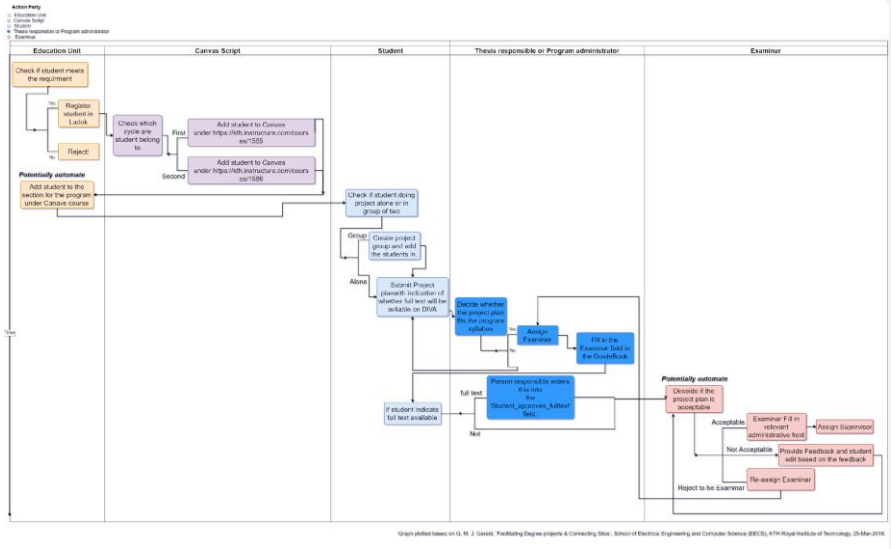

Figure 2-1: Workflow for Degree Project

At the completion of the degree project students submit a draft thesis and present it in a seminar (referred to as a thesis defense or presentation). These presentations are announced via Calendar events in Polopoly. Following the public presentation the thesis is typically revised and the final version submitted via Canvas to the examiner for evaluation. If the thesis is accepted then a grade is assigned by the examiner and the thesis should be entered into DiVA with a unique document number (at KTH this is called a TRITA number). The information in the thesis will be used to generate the front and back cover pages of the thesis via the Book Cover Generator (see Section 2.13). This step can be potentially automated with Data mining process as Figure 2 showed [3, 4]..

2.2 Learning Management System

A Learning Management System (LMS) is an information technology (IT) solution that provides support for administration, documentation, tracking, reporting, and delivery of educational courses or training programs[1]. At KTH , an LMS is used to deliver course material, share documents, manage assessments and facilitate communication between students, faculty, and other education staff [5].

2.3 Canvas Platform

Canvas is a Cloud-based and user-focused LMS developed by Instructure, Inc.[*] and widely used in higher education instituions worldwide [6]. KTH adopted Canvas as their LMS starting in period 1 of 2017 [5]. Before Canvas, KTH used a variety of different systems, such as Bilda, KTH Social, and Daisy as LMSs [5]. Via Canvas, student can find course material and guides provided by their teachers; keep track, of courses, news, and events via the announcement board; keep in touch with teachers and fellow students through discussion forums and instant messaging; hand in assignments; and take course quizzes and exams [7].

Canvas supports a Restful API for external applications to access and modify data in the Canvas database [8]. The Connecting Silo project will achieve its goal through a combination of GET/POST requests via the Canvas API. This API can be accessed via an OAuth2 Authentication Token that is generated for each user on request. When using this token the application has all of the same permissions as associated with the user's Canvas account. Using this API, it is possible to access and move data between Canvas and other programs.

In addition to the Canvas Restful API, Canvas also implements a Content Management System (CMS). Via the Canvas web user interface, a user can create content. For example, a user could create a survey that can be used as form (to replace the existing thesis application form) to collect data from the student. Subsequently this information could be used to separate the students into additional sections (beyond the initial section based upon the student's education program). In the case of a degree project, each student is added to a section based upon the student's Examiner and Supervisor. An example of a student in several such sections is shown in Figure 2-2. The use of sections enables an examiner, supervisor, or program administrator to easily see a view of the gradebook that shows only their students.

[*] In Europe Instructure Global Ltd.

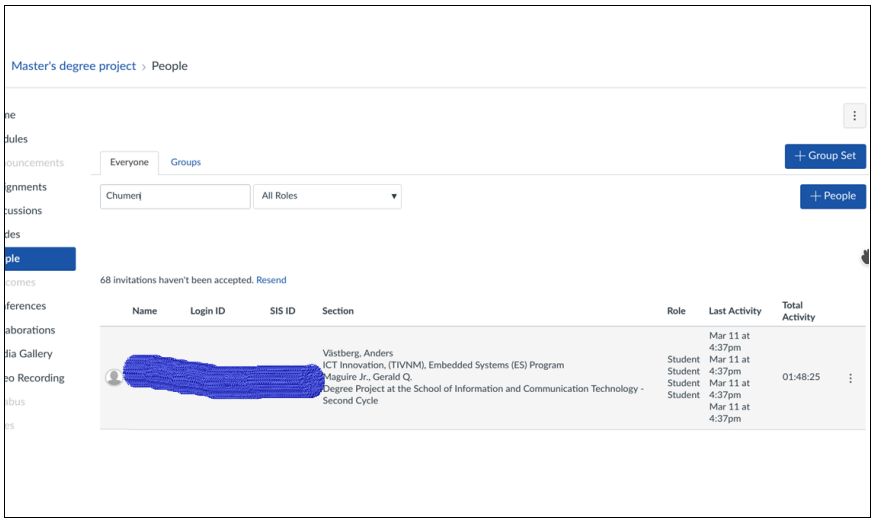

Figure 2-2: Example of a student placed in several sections

2.4 DiVA Platform

DiVA (Digitala vetenskapliga arkivet) - Academic Archive Online, is a publishing system for research and student theses and a digital archive for long-term preservation of publications [9]. Currently, there are around 47 different universities in Sweden using DiVA for publication registration [9]. At KTH, DiVA has been used to register researchers' publications [10]. These publication include doctoral dissertations, licentiate theses, reports, students theses, and other research publication where at least one other is from KTH. The purpose of the DiVA platform is to make publication visible and accessible via the DiVA platform [10]. DiVA also makes publication accessible by exchanging data with other database services, such as the Swedish national publication database (SwePub), Google Search Engine, and Google Scholar [10].

Unfortunately, the DiVA API is not accessible to the project team. This issue will be solved by generating MODS XML import file for subsequent processing as DiVA supports MODS file for modifying or creating records of publications. This file format will be described in Section 2.8. Additionally, there may be a possibility to insert records into the DiVA database using an existing API, as DiVA is based upon a Fedora Repository.

Kommenterad [m22]: See https://smartech.gatech.edu/bitstream/handle/1853/28430/139-648-1-PB.pdf

Another approach that project team might take on for DiVA API issue is to use Selenium Browser Automation plug in that describe in chapter 2.15. This approach will focus on autofill the DiVA publication forum and submit it without human attend during process. Eventually team will choose on one reliable approach and one approach that based on experimental but less effort for the user of the program.

2.5 Polopoly

Polopoly is a content management system at KTH which makes edit management and content publishing on the KTH website more natural [11]. Polopoly also has a function that can be used to publish a calendar event into specific calendar inside a specific user group. The event creation function in Polopoly is a useful function when it comes to announcements for thesis presentations. As described previously, the project should create a means to automatically create a calendar event in Polopoly based on the information extracted from the PDF version of a thesis draft in Canvas and the place and time for the oral presentation.

Kommenterad [QL23]: This chapter might need to re-wrote depends on the approch that we take for calendar system

2.6 Canvas Gradebook

The Canvas gradebook contains information about students submission for assignments, their grade for each assignment, and other information. In conjunction with the degree project course, the Canvas gradebook has been extended with additional custom fields to store information relevant to a degree project. Figure 2-3 shows an example of the custom columns that have been added to a gradebook for the 2nd cycle Master's degree project course. In this case the custom columns are those shown to the right of the column "Student Name". Some of the values stored in these columns will be used later to generate the cover for the thesis and to record the document number (TRITA number) assigned to the thesis, the DiVA URN that is assigned once the DiVA system has accepted a thesis, information about whether the student has given their permission for the full text to be published via DiVA, etc.

Figure 2-3: Example of the custom columns that have been added to a gradebook for the 2nd cycle Master's degree project course.

2.7 Speedgrader

Canvas Speedgrader allows a user to view and grade student assignments in one place with a simple point scale and complicated rubric [12]. Speedgrader allows enables the teacher to grade the submission and either directly add markup on the submission or using an external tool to do markup after downloading the submission as a file. Additionally, the user can easily provide comments to the student using Speedgrader. Speedgrader allows the user to grade, track, and provide feedback on the assignment in the Canvas without requiring additional tools [12]. In this project, Speedgrader will be used as an interface by examiner for entering feedback on a student's submission.

2.8 MODS

Metadata Object Description Schema (MODS) is a schema for bibliographic elements that can be used for several different purposes, especially in a library system [13]. The MODS schema language is encoded using the Extensible Markup (XML) format. MDS files can be generated by various tools to provide mandatory input data [14]. DiVA supports MODS as one of its publication import formats, thus enabling the automation of the workflow from other systems to DiVA.

2.9 Data Mining

Data mining utilizes a computer algorithm to discover patterns in large data sets by using machine learning, statistics, and database systems [15].

As depicted in **Figure 2-4**, the data mining process can be categorized into four steps: data sources, data exploration/gathering, modeling, and deploying models[16]. Data sources are the source files that are to be mined. The data sources will be minded by extracting specific data from the sources [16]. For the Connecting

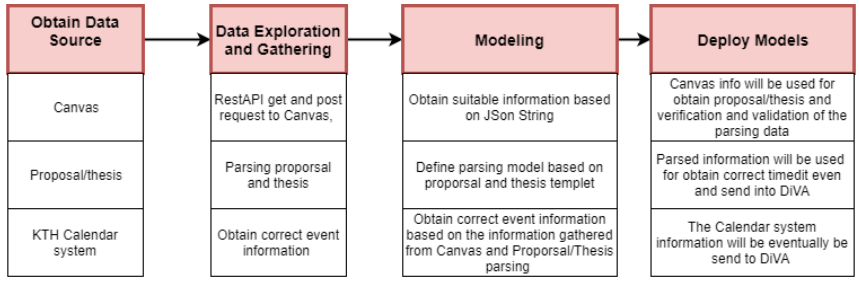

Diagram based on: MargaretRouse, 'What is data mining? - Definition from Whatis.com', *SearchSQLServer*,01-Dec-2008.[Online]. Available:http://searchsqlserver.techtarget.com/definition/data-mining. [Accessed:01-Apr-2018] [16]

Figure 2-4: The process of Data Mining based on reference [16]

Silo project, the data source is the PDF version of a thesis proposal or the thesis itself*. By means of data exploration/gathering, the pattern in the data sources is derived and the resulting extracted data transformed into a human or machine readable and-process able format. The data exploration/gathering method that Connecting Silo project will use is PDF parsing. The developer needs to specify a model to use in the data mining process. The model can be a pattern in the data, specific keywords to look for, etc. Finally, the deploy model will be created by the four steps in the data mining process and the extracted data will be used as input to other software [16].

* As an extension it should also be possible to mine DOCX files as they are well structured XML files.

The data that is mined from the data sources has to remain consistent. In the Connecting Silo project, this data consistency will exploit an ORCID (Open Researcher and Contributor ID) [17], KTH ID, or other identification system. ORCID is a persistent digital identifier that is used to identify authors and contributors [17]. KTHID is a digital identification that is used inside KTH for students, faculty, and administration staff [18]. In the Connecting Silo project, KTHID will be used to ensure data consistency regarding the examiner, supervisor, and student. The ORCID identifier will also be added to the records when known.

Kommenterad [gqmjr24]: Note that you cannot use someone else's figure without the copyright owner's permission.

2.10 PDF Parsing

PDF is a document format that is maintained by the International Organization for Standardization (ISO). ODF is a widely used standard for documents. Moreover, PDF is independent of the underlying operating system [19]. The PDF format supports links, buttons, form fields, audio, video, and business logic in a document's file [19]. To extract information from a PDF file, PDF Parsing is done to parse and analyze PDF documents [20]. PDF Parsing tools can extract the desired raw data from PDF documents and in some cases, even extract data from a damaged PDF file. PDF Parsing is one means of data mining. The information that is extracted by a PDF Parsing tool can be used by other systems.

Regarding the tools that will be used for parsing PDF - some prior work have been done by Elias Kunnas with his PDF Parsing tool called pdfssa4met [21]. Based upon this tool Gerald Q. Maguire Jr. wrote a program called kthextract to extract data from a thesis proposal or the thesis itself, Both tools are written in Python. The Connecting Silo project will implementation its tools based on pdfssa4met and kthextract. The project will further develop pdfssa4met and kthextract.

2.11 Pdfssa4met

Pdfssa4met is a python open source project by Elias Kunnas from Finland that aims to provide metadata extraction and tagging based on structural and syntactic analysis of content in XML [21]. The project is based on the the pdf2xml project, which is a tool to convert a PDF file into an XML file. After converting the PDF file to an XML file, pdfssa4met can be used to search in the XML for a specific keyword or XML tag. For example, to look for a reference (reference.py), the program will first look for the keyword "Reference" and then for the tag <reference></reference> [21].

2.12 kthextract

Based on Pdfssa4met, Dr Gerald Q Maguire developed kthextract.py to extract certain information from a Bachelor or Master's thesis. The project is based on python 2.x.x. The data mining part of the Connecting Silo project will be based on

kthextract project and will further develop this program to achieve the goal of the Connecting Silo project.

2.13 KTH Book Cover Generator

KTH's Book Cover Generator is used to generate the cover pages of a Bachelor or Master's thesis. The following information is needed to use the tool [22]:

- Cycle and number of credits of the degree project
- Degree
- Main field or subject of education degree
- Title
- Subtitle
- Author(s)
- (Optional) Image to be used as the front page
- School at KTH where the degree project was examined
- Year, and
- TRITA number (as a unique document number).
- The first and second cycle thesis do not require an ISRN and ISSN number

Kommenterad [gqmjr25]: Note that 1st and 2nd cycle theses do not have an ISSN or ISRN.

Most of the information that is required by the Book Cover Generator can be obtained from thesis itself, from Canvas, and from the KTH user database. Eventually, the front and back cover will be combined with the approved thesis via an existing PDF modification tool (such as PyPDF2).

2.14 PyPDF2

PyPDF2 is a python coded PDF toolkit developed from the pyPdf project. It is currently maintained by PHaseit[23]. PyPDF2 can create. extract, edit, merge, and encrypt & decrypt specific data from one or more PDF files [23p. 2]. PyPDF2 will be used to add the cover pages created by Book Cover Generator to the thesis. More specificially, PyPDF2 will be used to merge the thesis content with the front and back cover pages.

2.15 Python Selenium

Selenium is a browser automation plug in for python that aim for simplify the work flow for developer who have large amount of repetitive task on certain website[24]. For example, Selenium can auto fill the from on the website base on certain data set, set the profile for the browser and clicking a button or link on the website. To be able to use Selenium, developer should specify the path for the browser driver that one decide to use. For example, Firefox is using Geckodriver[25] and Chrome is using ChromeDriver[24]. For Connecting Silo project, the firefox has been chosen as default browser and geckodriver will be used as web driver for selenium to automate the thesis process.

2.16 Python Package Manager (pip and conda)

Python Package Manager is a mechanism to import packages into the python library[26]. Most well-known package manager are pip and conda. Since python has two widely used version at moment, depends on the version that developer used in their code the Python Package Manage might varies. For example, in pip, if one use python 3 .x.x for development, the developer should use pip3 instead of pip. The same goes for conda. For Connecting Silo project, python 2.7.14 are used for kthextract and python 3.6.5 are used for Canvas module but only Canvas module are frequently using external package. For kthextract only two external packages are used.

2.17 KTH API:er

KTH (Royal Institute of Technology Sweden) API(Application Programming Interface):er give public access to KTH common systems. With the help of KTH API:er one can obtain the results and registration, schema, course and program catalogue, KTH social, KTH profiles, WebTex(mathematics expression for web publishing), KTH directory, KTH web publish system and KTH places. In Connecting Silo project, the KTH API:er profile function has been used to obtain detail information of the student, examiner and academic supervisor[27].

2.18 Related work

The project is built on few relative work that has been done previously. The first one is kthextract that developed by Dr Gerald Q Maguire. Kthextract provide general function to extract fixed paged content and flexible page content. Based on kthextract,the parsing module will adapt to the EECS (School of Electrical Engineering and Computer Science) 2018 templet with few more function that Connecting Silo project required.

Second related work is the Canvas module that based on Dr Gerald Q Maguire previous work. The Dr Gerald Q Maguire version of Canvas module export an Excel format form that based on python pandas plug in[28]. The excel form include different categories of information from Canvas LMS(Learning Management System). Since Connecting Silo project do not need all the information in Excel format file, the standard output will be modified, and the output data will be selected.

Except Canvas token, the download submission function require KTH (Royal Institute of Technology) ID login. The reason is KTH website has a redirect function and the redirect function has extract 1 to 2 step redirections. To be able to download the submitted proposal. the Connecting Silo project used the python selenium web browser automation plug in. This plug in is not developed by the Connecting Silo project team as chapter 2.16 described.

2.19 Reliability Analysis

The Canvas module is based on RestAPI and most of the function is grabbing information based on provide input. The same combination of KTH(Royal Institute of Technology Sweden) ID, student ID, assignment id and course ID are always binding with the same output. Unless the administration department in KTH cause a registration mistake that made wrong binding between different IDs , the Canvas module do not exist big reliability issue.

The reliability issue can occur in parsing module due to the parsing system is based on identify the format of the text(for example bold, font and size) , position of the text (for example page, block and line) and the content of the text. This system will cause not cause big reliable issue as soon as student follow the proposal and thesis report templet that provided by KTH examiners, but issue can happen since developer cannot control user activity. At same time, the exception can exist in the model that define by the parsing module. For example, if the user need second content when two same content with same font exist in the same line. The parsing module will provide the first text as answer. In this case the output is not reliable.

Unit Test

Test Case	Requirement to Pass	Status and reason
Canvas Module		
Extract the Correct Attachment	The document that extract from the zip package which download from canvas fits the correct assignment and course code	
The zip file is complete	The zip file that extract from canvas contain assignments from all the student that has submit the assignment	
The student detail information is correct	The student detail information that extract from KTH Profile API:er is correct without modification	
The student list is correct	The student list that extract before the process module is correct. No extra student name is included.	

The link is usable	The link that Canvas Module extract from Canvas is downloadable. Downloadable mean it is not out of date, and no error report from the program in download stage	
Process Module		
The pdf file is accurately converted to XML	The pdf is successfully read into the system. The XML that converted is correct. (if this requirement failed, the problem is at external module pdfss4met and pdf2xml)	
The fixed page parsing is correct	The parsing result from fixed page parsing function is correct with out extra or less information. All the flexible parsing goal has been reached in current session	
The flexible page parsing is correct	The parsing result from flexible page parsing function is correct without extra or less information. All the flexible page parsing goal has been reached in current session	
The system is parsing the correct pdf file in correct stage	The pdf that is parsing belong to one of the students in the student list. The proposal is parsing in the proposal phase.	

	The thesis is parsing in the thesis phase.	
DiVA Module		

Integration Test

Test Case	Requirement to Pass	Status and reason
The DiVA from can be filled in correctly	The information that output after Canvas and Process module is sufficient to filled out all the mandatory field in DiVA publication form All the information that filled into the field are accurate	
The system is finished correctly	The program will finish without error report in all the module.	

2.20 Validity Analysis

With the fact the parsing module has reliability issue, the Connecting Silo project will also extract as much student, examiner and academic supervisor information from canvas as possible. The parsing module will compare the similar field to validate the document it is parsing is correct and if other parsing result is reasonable. If there is a mistake occurred during validation process, the system will either trigger

an error report from parsing module or choose the canvas output as result. The decision of which approach the Connecting Silo system will choose is based on the stage that error occurred. The extra information that parsed from canvas will be inject into the output of current parsing session as an extra field.

Before the project deploy, there will be two round of test session. During the test session, the test program will be deployed to the KTH (Royal Institute of Technology Sweden) administration stuff and collect feedback from the administration stuff. The feedback will include accuracy of the output data, user feedback (if the program makes the user uncomfortable) and sustainable development feedback (if they willing to use this program in their future work? In finically, environmentally and time-consuming point of view, how much resource does the program save?)

TODO: write a paragraph of feedback report after the report has been made. Also insert a figure about how the report is looked like.

Figure 2-5: Feedback report templet

2.21 Discussion

3 Conclusions and Future work

<<Add text to introduce the subsections of this chapter.>>

3.1 Conclusions

Describe the conclusions (reflect on the whole introduction given in Chapter 1).

Discuss the positive effects and the drawbacks.

Describe the evaluation of the results of the degree project.

Did you meet your goals?

What insights have you gained?

What suggestions can you give to others working in this area?

If you had it to do again, what would you have done differently?

3.2 Limitations

What did you find that limited your efforts? What are the limitations of your results?

3.3 Future work

Describe valid future work that you or someone else could or should do.

Consider: What you have left undone? What are the next obvious things to be done? What hints can you give to the next person who is going to follow up on your work?

3.4 Reflections

What are the relevant economic, social, environmental, and ethical aspects of your work?

References

<< Let Zotero or other tool fill this in for you. I suggest an extended version of the IEEE style – to include URLs, DOIs, ISBNs, etc. – to make it easier for your reader to find them. This will make life easier for your opponents and examiner.>>

Kommenterad [gqmjr26]: IEEE Editorial Style Manual: https://www.ieee.org/documents/style_manual.pdf

[1] Ryann K. Ellis, 'Learning Managament Systems', *Alex. VI Am. Soc. Train. Dev. ASTD*, 2009.

[2] D. Kuhlman, *A Python Book: Beginning Python, Advanced Python, and Python Exercises*. Platypus Global Media, 2011, ISBN: 978-0-9842212-3-3 [Online]. Available: https://books.google.se/books?id=1FL-ygAACAAJ

[3] 'Degree project | Bachelor's Programme in Information and Communication Technology (TCOMK, 180 cr) | KTH'. [Online]. Available: https://www.kth.se/social/program/TCOMK/page/degree-project-5/. [Accessed: 30-Mar-2018]

[4] 'Degree project process | Bachelor's Programme in Information and Communication Technology (TCOMK, 180 cr) | KTH'. [Online]. Available: https://www.kth.se/social/program/TCOMK/page/routine-for-degree-project/. [Accessed: 30-Mar-2018]

[5] 'Learning Management Systems | KTH'. [Online]. Available: https://www.kth.se/en/student/kth-it-support/learning-platforms/learning-management-systems-1.517117. [Accessed: 28-Mar-2018]

[6] Instructure Global Ltd., 'Canvas by Instructure', 05-Apr-2018. [Online]. Available: https://www.canvaslms.eu/. [Accessed: 28-Mar-2018]

[7] 'Canvas | KTH'. [Online]. Available: https://www.kth.se/en/student/kth-it-support/learning-platforms/canvas/canvas-1.784659. [Accessed: 28-Mar-2018]

[8] 'Canvas LMS REST API Documentation'. [Online]. Available: https://canvas.instructure.com/doc/api/index.html. [Accessed: 28-Mar-2018]

[9] '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?dswid=-1313. [Accessed: 28-Mar-2018]

[10] 'About DiVA | KTH'. [Online]. Available: https://www.kth.se/en/kthb/publicering/kths-publikationsdat/om-diva-1.569302. [Accessed: 28-Mar-2018]

[11] 'User's guide to Polopoly | KTH'. [Online]. Available: https://intra.kth.se/en/administration/kommunikation/webbpublicering/polopoly/manual/att-jobba-med-polopoly-1.8432. [Accessed: 28-Mar-2018]

[12] 'What is SpeedGrader? | Canvas Community'. [Online]. Available: https://community.canvaslms.com/docs/DOC-10712. [Accessed: 28-Mar-2018]

[13] 'Metadata Object Description Schema: MODS (Library of Congress Standards)'. [Online]. Available: http://www.loc.gov/standards/mods/. [Accessed: 28-Mar-2018]

[14] 'MODS: Uses and Features (Metadata Object Description Schema: MODS)'. [Online]. Available: https://www.loc.gov/standards/mods/mods-overview.html. [Accessed: 28-Mar-2018]

[15] Soumen Chakrabarti, Martin Ester, Usama Fayyad, Johannes Gehrke, Jiawei Han, Shinichi Morishita, Gregory Piatetsky-Shapir, and Wei Wang, 'Data Mining Curriculum: A Proposal (Version 1.0)', 30-Apr-2006 [Online]. Available: http://www.kdd.org/curriculum/index.html. [Accessed: 05-Apr-2018]
[16] Margaret Rouse, 'What is data mining? - Definition from WhatIs.com', *SearchSQLServer*, 01-Dec-2008. [Online]. Available: http://searchsqlserver.techtarget.com/definition/data-mining. [Accessed: 01-Apr-2018]
[17] 'ORCID'. [Online]. Available: https://orcid.org/. [Accessed: 30-Mar-2018]
[18] KTH, Department of Leaerning, Language and Communication Unit, 'Frequently Asked Questions', 28-Apr-2014. [Online]. Available: https://www.kth.se/en/ece/avdelningen-for-larande/sprak-och-kommunikation/verksamhet/tandem/frequently-asked-questions-1.378107. [Accessed: 05-Apr-2018]
[19] 'What is PDF? Adobe Portable Document Format | Adobe Acrobat DC'. [Online]. Available: https://acrobat.adobe.com/us/en/acrobat/about-adobe-pdf.html. [Accessed: 28-Mar-2018]
[20] 'InfoSec Handlers Diary Blog - PDF Babushka', *SANS Internet Storm Center*. [Online]. Available: https://isc.sans.edu/diary.html. [Accessed: 28-Mar-2018]
[21] 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: 01-Apr-2018]
[22] 'KTH Book Cover Generator'. [Online]. Available: https://intra.kth.se/kth-cover?l=en. [Accessed: 30-Mar-2018]
[23] 'PyPDF2 Documentation — PyPDF2 1.26.0 documentation'. [Online]. Available: https://pythonhosted.org/PyPDF2/. [Accessed: 30-Mar-2018]
[24] 'Selenium - Web Browser Automation'. [Online]. Available: https://www.seleniumhq.org/. [Accessed: 20-Apr-2018]
[25] *geckodriver: WebDriver <-> Marionette proxy*. Mozilla, 2018 [Online]. Available: https://github.com/mozilla/geckodriver. [Accessed: 20-Apr-2018]
[26] 'Installing Packages — Python Packaging User Guide'. [Online]. Available: https://packaging.python.org/tutorials/installing-packages/. [Accessed: 20-Apr-2018]
[27] 'Use data from KTH | KTH'. [Online]. Available: https://www.kth.se/en/api/anvand-data-fran-kth-1.57059. [Accessed: 22-Apr-2018]
[28] 'Python Data Analysis Library — pandas: Python Data Analysis Library'. [Online]. Available: https://pandas.pydata.org/. [Accessed: 22-Apr-2018]

Appendix A: xxx

Appendix B: Detailed results

TRITA-ICT-EX-2017:XX