DOCUMENT MANAGEMENT SYSTEMS IN HIGHER EDUCATION[1]

Gábor Percze[1], Nikolett Menyhárt[2], Gábor Kusper[3]
R&D Consultant[1], Demonstrator[2], Associate Professor[3]
UniWeb Plusz Ltd.[1,3], Eszterházy Károly College[2]

Introduction

„Information worth as much as it is used of it" – goes the quote I read the other day. Unfortunately, I can't mark its author. Among the main the properties of information is its value. This value can be social, economic, productive, scientific or even personal. Naturally, as it has a value, it can generate profit.

The industry based on information is advancing by leaps and bounds. Tough this statement is commonplace, it still has much truth in it. One characteristic of today's devices is the fact that a cheap smartphone has better data storage, computational and data traffic capabilities than a server machine produced a few decades before. The devices' speed of advancement is exponential, it is not possible to foretell what will happen in the next 5 or 10 years. Learning from the lapses of our predecessors, we are making statements like the below carefully. "Computer science has reached its limit." is a quote from János Neumann from 1939. It is clear that even the greatest minds may lapse. The following quote is originated to Bill Gates, tough he denies it: „640K should be enough for anybody!" This quote is from 1981. Considering the systems and devices today, Gates – if he really said it - is rightfully twitted with it.

In these days, there is an enormous amount of data on the worldwide web. According to the records [1], there are approximately 4,58 billion indexed websites on the internet. All of these sites are full of different information. If we would want to download the whole internet to a single device, it would require something with 672 Exabyte (≈672.000.000.000 GB!) storage capacity [2]. In 2013, the internet's data traffic has reached 43 639 Petabytes.

This mass of data has been created together by people, devices and sensors. To handle such a huge amount of data, we definitely need data-handling systems that help with the management and version control of documents as that task has clearly overgrown human capabilities.

In today's world, we have much more storage space than 640K. Thanks to the advanced devices of informatics, the flow of information is no longer limited by the different gadgets and the speed of data transmission. The next task to be solved is to handle the mass of data we have. When in need of information, we naturally turn to the internet. The issue is to ask the proper question to the device and even if the question is correct, does the answer know we asked about it?

In most of the cases, it does not. With the spread of the document management systems, the situation seems to improve, but the ratio of indexed, categorized data is still low.

In this chapter, we use different phrases. To help the understanding and avoid misunderstanding, we clearly define these phrases below.

[1] Supported by the Research and Technology Innovation Fund in the project PIAC_13-1-2013-0268 SCUNL.

DOI: 10.26649/musci.2015.091

Definition (digital space): The whole of the digitally stored data is called the digital space.

Definition (label): The whole of the data's parameters that can be used to unambiguously categorize the data is called the label.

Definition (Active data): The data in the digital space that has labels is called active data.

Definition (Passive data): The data in the digital space that has no labels is called passive data.

There is much information in today's digital space that are not properly analyzed and labeled, so they are hovering as passive data in the digital world. We can't say for certain that they will ever become active building blocks, but it's sure that their activity is not enough to list them as exemplary citizens of the digital world.

So the task is to make the passive data active!

The "data-hunger" of higher education

One characteristic feature of the higher education is the amount of data created that is stored in different locations. For example, in the University of Debrecen, the library data is stored on a different server than the data of the academic department. On most documents during academic work, more people are working on and it is available from multiple sources.

Looking at it from student's or researcher's point of view, we can state that the freshness of knowledge depends on the freshness of the data. One typical demand of today's generation is to get accurate, fresh information in the blink of an eye with minimum effort. The classic library and other general tools of categorization are not capable of providing this.

Thanks to the rapid flow of information and application ensuring active online social life, there is a demand for sharing common knowledge and for an interface where common knowledge can be shared. For example, if the programming lesson is cancelled, there will be a huge load of facebook posts about it. As a consequence, students are no longer hanging around the lecture room waiting for the teacher.

The outcome of the mentioned characteristics is the need for document-handling systems that can satisfy these needs. If we take a closer look on data, we can see that it has properties. A thesis, that is a characteristic data in higher education, has a title, and author and contains references. If we use the current searchers, we can look for, for example, a title. But there are some questions that can be raised in such circumstances. One such question is: where to look for it? In the library? In the system that handles thesis? Is it sure we will find it? Do we know if our search is good? Can we phrase what we are looking for?

Imagine the situation when in a document-management system in higher education, all data is active, all data knows if it had been searched for or if it is not the latest version.

If this would be the reality, our productivity could be multiplied. We would always find what we are looking for. We would even find things that we didn't know we are in need of them!

The document management systems used in today's higher education

Alfresco ECM

One well-known document management system is the Alfresco system. Its main purpose is not to serve the education, but the handling of corporate documents. Alfresco is an all-round document and content management system for large enterprises, which, in its functions, is in many ways the equivalent of the overcomplicated and expensive "classic" proprietorial ECM solutions, but altogether, thanks to its open-source approach and reworked architecture, it more cost-effective, and much more easy and cheap to integrate.

At its configuration, the main angle was to serve the needs of enterprises. These enterprises can be micro businesses, bigger multinational corporations or even governmental institution – it can deal with all of them. It corresponds to the recent expectations on document handling and uses the most up-to-date web technologies. At the same time, it is entirely open-sourced, which is a huge advantage for this kind of software, mainly for the dynamic development options, while meeting the expectations on services of huge institutions and corporations and providing the support for open-source systems. For the above, it's a leading solution for the document management issues of companies.

The services of Alfresco are not only available through the applications provided by the software, but it can be accessed from external systems using open and standardized institutions. The technological base is a central content storage space, which is capable of storing and handling documents in a reliable way. This central content storage space is called the Content Repository. The system offers external applications and services that are built on the central core, but anybody can create unique "shells" that can use the Content Repo.

One of the most important services of the Alfresco ECM is its ability to search. Simplified, this means searching in the content storage space, that is based on authorization, so the search results depend on the user's rights and accesses. There are more methods for searching. One such method is the SOLR, which is the search platform the Apache, allowing searches in big-sized systems with great performance. Another method is the Lucence, which is built on the Apache Lucenc, and it can search for metadata or can execute text only searches. The third searcher module is the path-based contextual navigation, based on Xpath specification. It also supports the SQL-based Query Language, that is standardized by the CMIS QL, OASIS – in the same time, supporting the Alfresco's query language. It also offers an option called "Alfresco Search Api", which can be used by developers to join in different search engines.

One of the base function of Alfresco is the support of workflows, for which it is using its own engine, the workflow engine, which can also be used to handle external workflows. This engine is responsible for executing the different workflows, for providing the multiple tasks and procedure definitions. This can't be accessed directly, only through the service called "workflow service", which provides a well-documented interface for reaching services. Thanks to the APIs, any application can be easily joint to the Alfresco ECM, while the developers also have the opportunity to create applications that utilizes the abilities of the workflow.

Handlind metadata is one of the essential functions of document and content management solutions. This is what differentiates them from the simple, file-based systems. The Alfresco ECM has advanced capabilities for handling metadata. It is capable of handling the metadata descriptors flexibly, according to the actual needs. Besides simply getting the metadata, the Alfresco ECM also provides different services. One of these is the automatic transformation of contents, which enables the handling of the different materials getting into the Repo.

The system, by default, provides a web-based interface for the end-users, but it also provides surfaces towards different boxed applications and custom development as well as towards different file system. In the same time, it supports standard web services.

Besides document management, the Alfresco ECM enables group work, receiving, storing, categorizing, and version handling. Blogs and wikis can be created and managed and the changes of documents can be followed with email or RSS messages. Its web appearance is fully customizable.

The Alfresco ECM is a totally open-sourced ECM solution, so when choosing the running environment, we have a much wider selection, than in the case of many proprietorial software. Moreover, the Alfresco ECM has been fully developed through Java technology, which ensured mobility in itself.

To run the Alfresco ECM environment, we need to have a Java application server, something to handle relational databases (supporting JDBC surface) and a platform that ensures the functioning of these components. In practice, an environment that fulfills the above mentioned requirements can be set up easily, so configuring a working Alfresco environment is simple and cost-effective.

Alfresco in education

Tough not entirely aimed to be used in education, it can be well-utalized in education as well. Such is the [3] initiative. Besides, the Corvinus University of Budapes is using a document management based on one such system. According to an initiative from 2010, Alfresco-based document-handling and storing system had been created on the Corvinus University. [**Hiba! A hivatkozási forrás nem található.**] Along with the Hungarian nstitutions, the University of Westminster is also using and Alfresco-based document-management system [5].

SharePoint

The SharePoint is an online group work application, that, through the use of different web applications, makes working together and cooperation easier for colleagues who are physically away from each other. The SharePoint is a product of Microsoft, so the applications found in SharePoint are compatible with the applications of the Microsoft Office product family.

SharePoint is capable of version control, filtering, handling documents, sending automatic email notifications, creating list views, content management, offers website functions, has search engine, supports user and user right handling, can be started without investment cost, customizable, expandable, offers constant availability. It also has voting machine, recycle bin and picture gallery functions.

It enables the creation of an environment where the users can easily work together with their co-workers. It is also capable of controlling different company procedures and has regular data saving functions.

Other initiatives in connection with Hungarian higher education – SharePoint 2010

The University of Kaposvár set itself an aim on creating a modern collaboration environment based on SharePoint 2010 with co-working and document surfaces with their unique image and with custom procedures.

The University's organizations run the handling of their committee meetings, presentations, the standardized storing of the establishment documentations, the handling of the request sheets for the internal services and the request procedures. One of the most important procedures in the University, the Purchase Procedure that supports the procedure regimes and supply circles has also been done through the system [6].

Google

Tough the system for document management offered by Google is not really created for huge enterprises, there has been more initiatives to use Google's cloud-based service. One such initiative is the agreement between the Kodolányi János College and Google. The entirely free Tutor Pack, offered by Google, has been created especially for elementary- and secondary schools and for higher education. It helps the cooperation, co-working, information sharing and communication between students and teachers, while reduces the administrative burdens and fees.

Among Google's applications, the web-based emailing service, Gmail, that provides a 15GB cloud storage space must be highlighted. Other important applications are the Google Calendar, which can be used to follow public events as well as personal, private appointments, the Google Talk, which is an instant messaging and file sending service also capable of VoIP calls and conferences, the Google Documents can be used to store documents and spreadsheets. These are all free, cloud-based solutions, capable of not only storing but sharing, the handling of user rights while allowing multiple users to work on the same document in the same time. Finally, Google Sites is used to create group websites, on which different information can be quickly collected and shared.

Google offers a different educational package for educational purposes. The next page contains references like the University of Notre Dame or the Sapienza University in Rome [7, 8].

KnowledgeTree document management system

The system called “KnowledgeTree” can be used to create the desired unified common knowledge that is easily accessible, manageable and expandable for all intended participants. The Knowledge Tree is a simple, web-based document management software that incorporates all the functionalities, positive attributes of the big producers’ products. What’s different is its being an open source application, thus having all the advantages offered by an open source application. The Knowledge Tree Community Edition, that has all the base function is free to download, use and modify. The KnowledgeTree has all the properties of the document management systems from big producers. Among the open source solutions, this is the most popular and most widely used document management system. The application is designed for team as well as for small- and medium-sized organizations. The Knowledge Tree is capable of working together with the Microsoft Office, Microsoft Windows and Linux systems. The document manager is a two-layer client-server application. The server side works on Windows and Linux operating systems.

It offers a web-based interface that is easy to use and is capable of supporting multi-language user surfaces. The common page is customizable to the individual’s needs (followed documents, issues in progress). It supports the WebDav access, which means that the document storage serves can be seen as a drive on the client’s side. It can handle many types of documents. The documents can be labelled, can receive zip files and can search in the label cloud. It can support different workflows that are based on tasks or teams. The workflows can be joined together the documents can be copied and moved. Notifications sent about the changes of a given document are also based on teams or tasks.

The access is realized through a browser-based client, but there are modules for OpenOffice and Microsoft Office. The mentioned WebDav can also be used to access the document storage space, and it also offers programmable APIs to develop other applications.

Trenoo

The Trenoo system had been created for educational institutions. It enables document management and also enables the evaluation of student works by the users. It also helps with education management and with the version control of commonly used documents. It can handle the portfolios of teachers and students and can upload all the common file types that are used for these portfolios. It also offers a mail system for the users. The data is stored in the cloud, so the work can be more efficient as the materials can be accessed from anywhere. The Trenoo also supports the different workflows and offers online administrational services. There are different financial and management services integrated to the application. Students can access the information stored about themselves and their financial transactions. The system can follow different contracts and notify the users when such contracts are about to expire and handle different due dates. It utilizes and automatic data deletion mechanism, that deletes documents that are more than 7 years old, thus sparing resources for the institution. Sure enough, it does not delete anything without notification, before executing the operation, it sends a notification to the respected users. It offers mobile access to the system, so the data and information stored in it are accessible anywhere. It helps the electronic administration.

The document management and social services of the Neptun system

The Neptun is the mostly prevalent system in Hungarian higher education that can be used to manage and administer scholastic procedures. The Neptun, besides of storing the students' official information and managing administrational tasks, is capable of storing different documents. Neptun's *Neptun Meet Street* module offers social services.

Neptun can handle different roles and complex user rights. It differentiates student and teacher modules in the scholastic and social systems. It created the opportunity for students and teachers to create and manage so called virtual spaces in connection with the school that can be used communicate through forums with their fellow students and with their teachers, add documents to this virtual space that can also be shared within the virtual space. The NMS (Neptun Meet Street) also enables the students and teachers to manage and administer their tasks in connection with the institution. Another NMS service is the option for the teachers to make their training materials (in connection with their different courses) public and students can share their thoughts on these published materials.

The social and education management module is linked. Tough it requires a separate authentication, it offers the same settings.

It also provides a service to send and receive messages and supports Outlook export, thus offering a complete mailing service for its users. It also defines virtual spaces that are interface for students and teachers to exchange thoughts and express their opinions on different subjects.

Document management is also supported. Documents can be uploaded for personal use or for publication – even with the members of a virtual space on a topic. Communities can be created with members of certain groups (students of a course), or new virtual spaces or groups can be created. News can be added to the virtual spaces. These can be placed and studied.

We can differentiate three kinds of virtual spaces: Subject, Course and Other. The documents that were uploaded to the virtual space can be edited on online mode. These common documents are available for the users they had been shared with.

Neptun is capable of playing and displaying electronic training materials, so it supports the pursuits of E-learning. This module can be accessed both from teacher's and from the student's web. It's important to mention that these training materials also support the SCORM E-learning standard, so old training materials are easily taken over while new ones are easily created. The framework is also capable of handling different versions of training materials and supporting multi-language training materials. The system can also create different statistics in connection with the training materials, handling the results of the students on the course and different tests can be created with its help. Blogs, message boards, polls can also run in Neptun. Similar to the products of Google, Neptun offers a calendar function and automatically displays the student's timetable there.

A personal document storage space can be assigned to the user account, where files and documents can be uploaded. We can assign other users to our own documents, thus sharing them with other people. Groups can also be assigned to the file. The documents can be downloaded in a pack and, according to user rights, they can be deleted or updated – and of course, these updates can be tracked.
The system can list the tasks assigned by teachers, so the submitted materials can be easily evaluated. The student is automatically notified about the result given.

How to collect passive data?

There are more methods to collect passive data. One method can be described by considering the data in a library. The problem with this is, it does not contain all the data. If we look for the weekly menu of a restaurant, won't find in the library the information that it'll be chicken soup with poppy-seed doughs. And the library is quite a static set of data. Even the best library's order of procedure is too complicated for a quick search. Hungary's largest library, the National Széchenyi Library has around 15000 printed books, more than 1200 manuscripts and hundreds of maps [9]. On the contrary, this is only a small fragment of the data available on the internet.
Another method is to create a resident program that collects the information it considers useful by keeping an eye on the data communication and then categorizes them. The realization of the above seems quite cumbersome and in the same time, it slows the data communication.
The base of the third method is a statement:
Statement: The digital data nascent on the university are available on websites connecting directly or indirectly to the university.
The data in the virtual space in connection with the university are thus available. Most of these data is passive data. The task is to activate these data. Starting from the meaning of "Big Data", we need to take notice of those data that are created by intelligent networks, users, corporations, members of digital spaces and other data-manufactures. These data are a huge amount. According to IBM, we create 2,5 exabit of data every day, which counts as 2,5 trillion bits. The biggest problem is, of many data, we don't even know it exists – but we store it nevertheless, it rests on the data storage of one of our devices. In many cases, these data has useful information that can be used to create value.
Statement: After activating a passive data, it van produce profit.

The goals of the development at the University of Debrecen

We started with the following statement at the beginning of the development:
Statement: The technology is mature enough to fill the flow of information with content.
Therefore, the goal is to activate the passive data. We would like to form a new document management system that satisfies the generational demands. The tool for this is to ask for help from the methods of data-mining after the data has been collected to select which data belongs to which category and put a label on it. The

current systems can only follow the versions and changes of a document. Our task is make functions that makes the system capable of categorizing.

During the development, we have encountered a question: where will the data receive the label?

The answer to this question is easier to understand if we imagine the data as a citizen of the digital space, who is travelling on the digital information highway. One terminus on this digital highway is the house of our digital citizen. Our digital citizen (hereafter: X) is just leaving for work. In this situation, X steps into the digital highway and is on his way to his workplace. When X gets to his workplace, he will surely be there for some time. After the day's work is done, X can decide to go home, but in the same time, his boss can ask him to travel somewhere to do work, so he will stay in a hotel today. Anyway, the example shows that X spends more time on some termini. The other statement is, X is surely using the digital information highway between two termini.

According to this, X appears on different termini during a set time period. Our task is to find these termini and evaluate the actual state of X than give him an appropriate label. We have nothing else to do with him.

If we make the answer a bit more accurate, we can state that a given passive data goes through more termini before reaching its destination, so these termini are suitable to give a label to the data.

Analyzing the answer, we came to the conclusion that we need the following tools to analyze the data:

- **Digital highway** – we already have this, let our digital information highway be the network of a university.
- **Termini** – the termini are also available in the network. Think of the fact that the digital data can only be considered data if it is stored on a digital data storage device of a kind.
- **Labelling machine** – we don't have the labelling machine yet. We need to create it (more of them) and place them on the termini.
- **The Boss, which is a factor that causes the move of the data** – in the case of X, his boss had an effect on him that caused X to move on a different direction on the information highway. Therefore, in any case, we need an agent that influences the path of the data. This agent can be a move command executed by a user or the event of downloading a website.

Okay. But why is it good for the data to have a label?

If we have a label on the data, we can identify whom does it belong to. It follows, we can pass it to the one requires it. Moreover, the active data is kind enough to look for its own owner.

After reviewing the process we can see how the activation of the data helps the production of profit. Let's take the example of a research: while doing the research, we might not even know how much passive data is flowing into our system, which are created by devices. In many cases, it can happen that we have passive data that is useful information for us.

During the development, we take into account that the primary device for the current generation is the smartphone. Thus, the system needs to adapt to the generational needs. Thanks to this, we need to create applications that are accessible through mobile clients for the users.

Statement: The needs of the current generation can mostly be satisfied by mobile applications, the mobile phone has become the primary IT device.

If we take the two statements, we can draw the following conclusion: the document management system can be tested in laboratory circumstances in the "mini digital space" provided by the university; it must be accessible by the users through the current primary IT device, the mobile phone; the data of the university, the appointed digital space, is already available from the websites in connection with the university.

Therefore, we need to re-define the data communication procedures. The technological world is clearly moving towards the mobile phones. According to the current trends, we need to create a system that follows the current trends, the used data flows. Naturally, when designing the system, we didn't start all from scratch. The design principle was the following: "Take the current, existing systems as a base, analyze them to find their errors and flaws, correct them and make the system adoptable."

Before the development procedure, we did stand for creating a solid base for it. So, we considered what kind of data occur in the appointed digital space, what are the primary IT devices, what form of communications are used in our everyday lives, what operating system we need to adept to.

Summary

The goal of this article is the review of our developments. During the research and development procedure, we have gathered many ideas that make us confident and assertive about the development itself. Our goal is to create the document management system of the future. We are lucky enough to work on a field of science that is in a constant state of advancement.

In our article, we reviewed the current state of the technology and the needs of the users. We discussed what systems are in use in Hungary and in the world around us. We took this as a base to introduce the direction of our development. We also introduced a few definitions and made statements that were unavoidable to take into account the requirements of the current age.

In the future, we would like to continue our persistent researches. Looking forward, even besides the current developments, there is a repository of challenges ahead; the besting of these is a difficult but gripping task. According to the experiences so far, our world is still full of excitement. Even with the best intentions for advancements, it can happen that our developments will become out-of-date, as we are competing on a field where the circumstances are constantly changing. Despite all of the above, our results so far are giving us a promising picture of the future.

Bibliography

[1] http://www.worldwidewebsize.com/

[2] http://www.factshunt.com/2014/01/total-number-of-websites-size-of.html

[3] http://www.alfresco.com/news/press-releases/alfrescos-content-management-and-workflow-system-chosen-mcgraw-hill-higher
[4] http://nws.niif.hu/ncd2011/docs/phu/004.pdf
[5] https://www.alfresco.com/customers/university-westminster
[6] http://www.avander.hu/post/2012/01/18/Dokumentumkezeles-a-Kaposvari-Egyetemen.aspx
[7] https://www.google.com/work/apps/education/
[8] https://www.google.com/work/apps/education/customers.html
[9] http://www.oszk.hu/