ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Semantic Analysis of Accreditation Document Using Ontology

Mark Cherwin L. Alejandria[1], Dr. Albert A. Vinluan[2]

DIT Student, School of Graduate Studies, AMA University, Project 8, Quezon City, Philippines[1]

Professor, School of Graduate Studies, AMA University, Project 8, Quezon City, Philippines[2]

ABSTRACT: This research aimed to develop a computing solution for the Accreditation Committees of Higher Educational Institutions in the Philippines that will help them in categorizing document for exhibit. The title "Semantic Analysis using Ontology for Document Analysis" focuses on the development of the domain ontology which includes document extraction and classifying document to its area and even sub-areas of the exhibits, developing an algorithm and evaluating its efficiency and evaluating the prototype by the respondents using ISO 9126. The researcher used evolutionary prototyping wherein the developers learned from the user a more accurate end products that allowed flexible design and development. The prototype was developed using different tools for extraction, especially Optical Character Recognition (OCR), ASP.NET as front-end and SQL Server for the database, as well as, structuring its own ontology. Unstructured files are basically the input for the system. A dashboard is also developed to monitor the number of documents in all areas of accreditation. The researcher conducted an acceptability test with the respondents which are the members of the Accreditation Committee who clearly emphasized that the developed system has all the features, functionality and modules that satisfied the overall requirements of the stake holder.The compute overall weighted mean is 4.76, which means that the respondents strongly agreed on the different criteria used for evaluating the software such as functionality, reliability, usability, maintainability and portability.

KEYWORDS: Algorithm, Dash board, ISO/IEC 9126, Ontology, Optical Character Recognition (OCR).

I. INTRODUCTION

Ontology usually provides a vocabulary describing a domain of interest and design of the meaning of terms in the vocabulary. Contingent on the precision of this specification, the notion of ontology comprises several data or conceptual models, e.g. classification, database schema, fully axiomatized theories. Ontologies then could be applied everywhere. They were viewed as the silver bullet for many applications such as database integration, peer-to-peer systems, e-commerce, semantic web service, social network. A document, in manual sense, is a form of information recorded on paper. However, in the advent of computer technology, a document can also be represented in an electronic form and stored in a computer as one or more files. The rapid advancement in digital technology has enabled significant changes in the ways companies handle electronic documents. Modern communication methods such as email, internet forms and digital video and sound files, accelerated business processes and gave the term "document" a new definition. Nowadays, a single document becomes a single digital file where the entire document or any individual parts can be treated as individual data items for automated processing. Such advancement has prompted many companies to rely heavily on digitized or electronic business documents. The existing practice of accreditation allows the institution to prepare and submit documents on processes, company policies, etc. which can be mapped manually to track compliance to accreditation standards. This process only provides a checklist tool that can be used to keep track of compliance requirements, where the requirements were manually indicated. To date, there are no tools that process submitted documents that automatically identify which area of the documentation conforms to a specific requirement in the standard or framework and to isolate which are partially accomplished and which are lacking. Multiplying these issues significantly further, many of the documents that needs to be presented from distinctive sources that may need to

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

be copied and distributed to different sections of the accreditation exhibit or may be required by other departments during their accreditation. Besides the accompanying cost in paper duplication, segregation issues may likewise exist in every department, like misfiling of documents, absence of office space because of the capacity of different papers, efficient searching techniques and some accessibility security problems that can only see, search and duplicate documents.

This study hypothesizes that the advancement in the field of computer-based text processing and digital technology, may be possible to capitalize on existing tools, resources and techniques and to apply it to the accreditation domain ontology with the goal of automatically determining level of compliance, as well as identifying gaps in compliance through automated processing of electronic documents.

One approach that can be considered to improve such systems was to utilize information extraction algorithms to automatically extract relevant information from electronic documents. In this study, the proponent's intention was to be able to extract information from documents and use it to generate a populated ontology of policy and map it against an ontology representing the different standards and frameworks of accrediting agencies. Generating the ontology of policy requires information to be extracted not only from a single document, but from a multitude of inter-related documents which serve as evidences of each of the criteria of accreditation. This means that application of information retrieval techniques was also necessary. The main purpose of information retrieval was to identify and select important documents, after which information extraction could then be used to extract information from the identified relevant documents.

There was at least one piece of information in every document. Furthermore, a group of similar documents also contains collective information. The process of extracting such information from text documents was often referred to as Text Mining. As its importance grew, the research on text mining developed impressively over the last decade, Text Mining tasks were extended to document retrieval, document summarization, entity extraction and modelling, information extraction and tracking, and sentiment analysis.

In the last few years, technology has advanced rapidly, enabling significant changes in the ways companies communicate with each other. Modern communication methods such as email, internet forms and digital video and sound files, accelerated business processes and gave the term "document" a new definition. Today, many companies rely heavily on digitized or electronic business documents.

Ontology is a knowledge repository in which concepts and terms are defined as well as relationships are between these concepts. It consists of axioms, relationships and set of concepts that describe a domain of interests and represent an agreed-upon conceptualization of the domains "real-world" setting. Implicit knowledge for humans is made explicit for computers by ontology.

The primary purpose of the study is to develop a system entitled "Semantic Analysis of Accreditation Document using Ontology" that can be used as a tool to have a convenient way of filing the necessary documents that were needed in compliance to accreditation standards.

- How to develop domain's ontology of terms and file extraction from any type of file to perform semantic analysis.
- How will the system classify the documents according to the area of exhibit based on the HEI accreditation and assessment requirements.
- How to design the dashboard that automatically monitors the accreditation related documents.
- How the respondents may perceive the level of acceptability the developed prototype with respect to chosen criteria under ISO/IEC 9126.

The general objective of the study was to develop a system that helped the Accreditation Committee of an Institution to manage and track the documents needed in order for them to apply for an Accreditation.

This research aimed to achieve the following specific objectives: To develop a solution utilizing domain's ontology of terms and file extraction from any type of file to perform semantic analysis.

- To implement classification technique in text document that was classified according to the given criteria.
- To develop a system that classify the documents according to the area of exhibit based on the HEI accreditation and assessment requirements.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

- To design the dashboard that will automatically monitors accreditation related documents.
- To evaluate the level of acceptability by the respondents through ISO/IEC 9126.

The beneficiaries of the proposed study were the following:

- The proposed system may advance of its information processing and retrieval.
- The system can minimize the time of the accreditation and the efforts of filing all the necessary documents needed in accreditation of the accreditation committee.
- By conducting this study, the researcher was improved his knowledge in different ways about information retrieval of any unstructured file and ontology engineering.
- This may serve as a guide for researchers to conduct the same kind of the study. They can apply the theories, knowledge and concepts to enhance their abilities or ideas to develop a new software.

II. RELATED WORK

The researcher examined books, journal and surfed various websites to get updated information needed in the study. Furthermore, the references elaborated the relationship and significance of the previous study with the present study which gave the researcher a better understanding on the development of software which was mainly the focused of this study. This helped the researcher acquire ideas on what to comprise in the proposed study and determine its capabilities and limitations.

Ontology was widely used as a mean to represent and share common concepts and knowledge from a domain or specialization such as knowledge representation, the knowledge within ontology must be able to evolve along with the recent changes and updates within the community practice. The knowledge extraction process uses ontology-based and pattern-based information extraction technique. Not only the extracted knowledge enriches the ontology, it also validates contradictory instance-related statements within the ontology that was no longer relevant to recent practices (Fudholi, 2016).

Ontology specifies rich semantic and shareable domain knowledge. The semantic richness provokes a broad study in ontology applications (Rahayu, 2016). Ontology has a chance to become a global community knowledge representation. As a knowledge representation, it was important to keep the knowledge within ontology to be consistently valid against the current practice. In order to achieve such validity, ontology needs to evolve along with the rapid change of knowledge.

This challenge was in line with the current challenge in Big Data, where data velocity was increasing (Armour, 2013).

Ontology-based File Extraction technique can be used to learn new knowledge from external sources and store the knowledge within ontology. Ontology-based File Extraction system populates and enriches ontology using extracted knowledge from natural language text. While being populated and enriched, ontology guides the extraction process to identify and extract related entities (Wimalasuriya, 2010).

The prototype was implemented as an automatic system, where the document source corpus can be searched automatically by using keywords that were extracted from the ontology components. The researcher used ontology based along with linguistic domain independent pattern-based information extraction technique, to extract related instances from recent related documents. The study was aimed to create a rule-based algorithm to derive the instance related statements.

The validation concept using recent knowledge seems to be similar to the concept of crowd sourcing system. Both concepts take recent knowledge to give some confidence in stating knowledge. Crowd sourcing enlists a crowd of humans to help solving a problem and needs to engage a community actively, which was time consuming and requires the appropriate financial and technical resources.

Text mining was nothing but the discovery of interesting knowledge in text documents. The interesting issue was how to guarantee the quality of discovered relevant features which was in the text documents for describing user preferences because of the large number of terms, patterns and noise. For text mining, there were basically two types of approaches; one was term based approach and another was phrase based approach. The term based method suffered with the

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

problem of polysemy and synonymy and phrase based approach suffered with low frequency occurrence. The phrase based approaches were better than the term based approaches. Pattern based approach was better than the term based and phrase based approach. The proposed process utilizes pattern method with the set of keywords, which was an innovative and real pattern discovery technique by which accreditation and assessment classification of different fields were done and more than 80 percent of documents were successfully identified (Chaudhari, 2015).

Text Mining was the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element was the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining was different from what were familiar with in web search. In search, the user was typically looking for something that was already known and has been written by someone else. The problem was pushing aside all the materials that currently not relevant to your needs in order to find the relevant information. In text mining, the goal was to discover unknown information, something that no one yet knows and so could not have yet written down.

Text mining has a solution to this problem by replacing the human reader with automatic systems Undeterred by the text explosion. It involves scrutinizing an ample Collection of documents to determine previously unknown Information. The information might be relationships or patterns that were buried in the document collection and which would otherwise be extremely difficult, if not impossible, to discover. Text mining can be used to analyze natural language documents about any subject, although much of the Interest at present is coming from the biological sciences. Originally, research in text categorization addressed the binary problem where a document was either relevant or not. Text mining comprises the use of techniques from areas such as information retrieval, natural language Processing, information extraction and data mining (Lade and Chaudhari, 2015).

Information Retrieval (IR) systems identify the documents in a collection which match a use's query. As text mining comprises applying very computationally exhaustive algorithms to large document collections, IR can speed up the analysis considerably by reducing the number of documents for analysis (Chaudhari and Lade, 2015).

Literature and studies both local and foreign were reviewed in this study that emphasized the significance on how to properly approach the study and the development of the system. The ideas and knowledge acquired by the researcher from reading the selected related literatures also corresponds to their own knowledge about the research.

In the research of Wimalasuriya and Dou (2010), Ontology-based information extraction: An introduction and a survey of current approaches, the researcher presented three key characteristics of OBIE systems that make OBIE systems different from general IE systems. The first of them was the process of natural language text documents. The inputs of OBIE systems were limited to unstructured or semi-structured documents. The next thing it encompass information extraction process, which was guided by ontology to extract things such as classes, properties and instances. The third characteristic presented by authors was a possibility to present the output using ontologies: OBIE systems must use ontology to represent the output. It was worth to notice that Information Extraction has employed various algorithms and methods for information retrieval. OBIE does not require a specific algorithm or method. It was based on ontology, which was used to support and guide algorithms for efficient and relevant information extraction.

In literature, the most frequently cited general architecture of OBIE system was proposed by Dou et al, (2010). In many cases OBIE systems do not include or exploit these components. Some elements from this general architecture were used to construct the procedure. Base on this, the proposal of a general architecture of OBIE system was presented. It was still developed and modified. It contains the additional modules: QAS selection support that was presented in a research entitled A Knowledge-based Approach to Question Answering System Selection (Konys, 2015), a tool supporting automatic ontology construction processes which was presented in A Tool Supporting Mining Based Approach Selection to Automatic Ontology Construction, and an approach for OBIE system selection and evaluation. The Protégé software was used as an ontology editor. It was worth to notice that OBIE system can be a part of a larger system. On input, a pre-processor component converts the text to a format that can be handled by the information extraction module. Information Extractor populates the Knowledge base and was queried by users. Output of the OBIE system contains the information extracted from the text. In addition, the output might also include links to text documents from which the information was extracted in the research Ontology-based information extraction: An introduction and a survey of current approaches in Journal of Information Science (Wimalasuriya, 2010).

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

The study started with accepting electronic documents that were stored as text files. The electronic version of the documents served as the input for the system. Each document was processed through the Ontology-based Information Extraction and fields would be filled in the system.

Selection of Knowledge Source started with the keywords generation and followed by document searching using the keywords. In the Knowledge Extraction process, it uses ontology-based and pattern-based information extraction technique. Not only the extracted knowledge enriches the ontology, it also validates contradictory instance-related statements within the ontology that was no longer relevant to recent practices. In the application, the instance extraction process and the statement extraction process can be done. Knowledge Enrichment and Validation enriches the ontology by adding new knowledge and validates the knowledge with the current knowledge which may result with knowledge alteration. In the present study, the researcher developed system that can automatically get the content or the main topic of a certain document to help the agencies compile all the documents in each category.

A conceptual framework is a group of concepts that defines and systematically organize a logical and sequential design created to outline possible courses of an action and to provide a less complicated structure based on the existing theory.

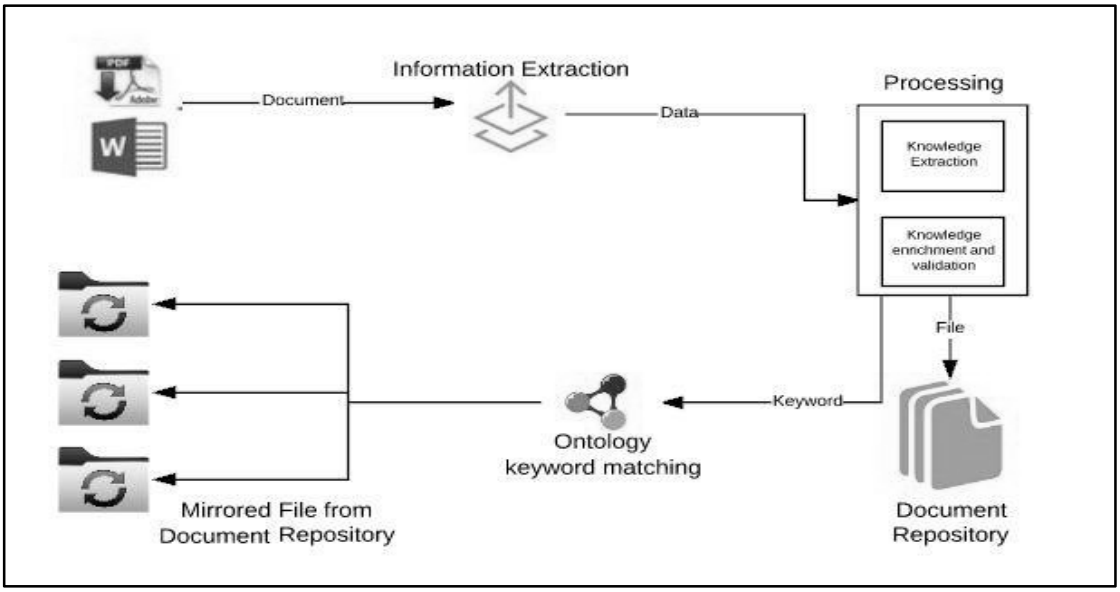

Figure 1: The Conceptual Framework

The conceptual framework has six steps, this would start by uploading the document or PDF file which is stored on the document repository which is a database and when it is collected in the repository, the information extraction begins.

During information extraction, pre-processing of the text occurs from the documents that have been extracted. Selection of knowledge source starts with the keywords generation and followed by document searching using the keywords. The keywords are used to match related keywords in the categories of the accreditation which were set by the accrediting agencies. The knowledge source searching involves the usage of identified keywords to find the matching keywords in the ontology.

In Knowledge extraction, it should have extracted and would identify relevant instances from knowledge source documents. In Knowledge enrichment and validation, the list of extracted instances and statements from the knowledge extraction phase enriches and validates the ontology in this last phase of pre-processing.

After the pre-processing, the extracted keywords will undergo ontology matching to determine keywords regarding the categories of the accreditation which are matched with the data file and store it in the repository. The information of keywords or ontology of terms are also being stored in the databases.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Data will be stored in the database, specially the information from the uploaded files such as content of the file, the name of the file, and the date it was uploaded.
From all the documents being uploaded, the system has a dashboard that will keep track and monitor the accreditation exhibits. The system will show the number of document per area and will alert the user if there are areas that are lacking documents.

III. METHODOLOGY

This section presents the research design, research questions and methods adopted in this study.

Research Design

Research is a rigorous, systematic investigation of a situation or problem to validate existing knowledge and generate new knowledge. Constructive Research is fundamentally done by numerous keeping in mind the end goal to discover new answers for any specific was sues. Productive exploration is the development, considering the current learning utilized as a part of novel routes with a couple of missing connections and base it upon theories, hypothesis, and case studies. The researcher used constructive research method for this study.
Constructive research is perhaps the most common computer science research method. This type of approach demands a form of validation that doesn't need to be quite as empirically based as in other types of research like exploratory research. Nevertheless, the conclusions must be objectively argued and defined. This may involve evaluating the "construct" being developed analytically against some predefined criteria or performing some benchmark tests with the prototype. The term "construct" is often used in this context to refer to the new contribution being developed. Construct can be a new theory, algorithm, model, software, or a framework.
In creating the prototype for the research "Semantic Analysis of Accreditation Document using Ontology" used a system development model in creating the prototype which had been discussed on this chapter. The researcher will have built an application in which all the digital documents in any type had been extracted to build the ontology of terms that had been used in data classification.
Appropriate algorithm in classification of document had been considered in the developing the prototype. The researcher looked for an algorithm wherein all the digital document had been classified according to the area of accreditation a specific digital document had been stored and categorized. The chosen algorithm had been evaluated in terms of its efficiency.
The design of the prototype addressed the problem in accreditation procedures, especially in organizing document for exhibit in terms of knowing if one of the areas of accreditation has an insufficient document and needed to comply. Finally, the researcher looked for a model on how to evaluate the effectiveness of the prototype.

System Development Model

One of the most appropriate software development life cycles is the Evolutionary Prototyping (EP); EP which refers to the activity of creating prototypes of software applications, for instance, incomplete versions of the software program being developed. It was used to visualize some component of the software to limit the gap of misunderstanding the customer requirements by the development team, as well as reducing the iterations that may occur. With this approach, the researcher presented the prototype to the stakeholders to obtain clarification and acceptance of the requirements before the design phase. As the prototype evolved, the researcher eventually set final requirements as agreed upon by the stakeholders until the final product was released.
In Evolutionary Prototyping, developers learned from customers for a more accurate end product and allowed for flexible design and development. Interaction with the prototype stimulates awareness of additional needed functionality of the system.
Since Evolutionary prototyping is a series of iterations, it is easy for customers to see some developments. In this case, gaining a positive impression is attainable from the customers because they can see how well the systems are improving to better handle customer service since the project is only considered complete once there is already the creation of the perfect system. The Evolutionary Prototyping software development model process begins with

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Requirement gathering and the iterations in its life cycle which are Analysis, Design, Coding, Integration testing and Maintenance.

The phases of Evolutionary Prototyping are:

Requirement Gathering. The first phase of the Evolutionary Prototyping life cycle is the Requirement Gathering, where stakeholders meet with the development team to engender user requisites. The researcher converted user stories into iterations that cover all parts of the functionality of features required. On this phase, requisites and information were accumulated meticulously and then analysed so that opportune allotment of time, cost and effort were done prior to the design and coding phase.

Analysis. The analysis phase is a process of collecting the gathered data; understand the processes involved, identifying problems and recommending feasible suggestions for improving the system. Business processes, gathering operational data, understanding the information flow, finding out bottlenecks and evolving solutions for overcoming the weaknesses of the system were involved in analysis phase to achieve the organizational goals.

The major objectives of the Analysis phase are to find answers such as:

1. What was being done?
2. How was it being done?
3. Who was doing it?
4. When he was doing it? Why was it being done?
5. How can it be improved?

The result of the process is a logical system design. Analysis is an iterative process that continues until a preferred and acceptable solution emerges.

Designing. The researcher divided large modules into smaller ones called spike solutions in designing phase. Several modules correspond to the stakeholder's requisite. These modules were considered in designing the repository. The graphical user interface was done regarding the requirement and per module. The system was designed as simple as possible at any given moment and extra complexity was removed as soon as it was discovered. Designing is considered the crucial phase in the development of a system. The logical system design arrived as a result of system analysis and was converted in physical system design. Input, output, databases and processing specifications were drawn up in detail. In designing phase, the programming language and the hardware and software platform were decided. Data structure, control processes, equipment source, workload and limitation of the system, interface, documentation, and training, procedures of using the system, and staffing requirements were also decided in designing phase.

Coding. This was the most consequential phase of the Evolutionary prototyping life cycle. Evolutionary prototyping gave priority to the authentic coding over all other tasks such as documentation to ascertain that the stakeholder would receive something substantial in value at the end of the day.

The system design needs to be implemented to make it a workable system. Coding is an important stage where the defined procedures and requirements are transformed into control specifications by the help of a computer language. A well written code reduces the maintenance and testing effort.

Integration and Testing. Before actually implementing the new system into operations, a test run of the system was done. Removing all the bugs, if any. Testing is an important phase of a successful system. After codifying the whole programs of the system, a testing should be developed and run on a given set of test data. The output of the test run should match the expected results by using the test data such as Program test and System Test.

Program Test happens when the programs have been coded and compiled and brought to working conditions, they must be individually tested with the prepared test data. All verification and validation be checked and any undesirable happening must be noted and debugged.

System Testing was done in actual data. The complete system was executed on the actual data at each stage of the execution, the results of the system was analysed. During the result analysis, outputs found that it did not match the expected output of the system. In such case, the errors in the programs were identified and were fixed and further tested for the expected output. All independent modules had been brought together and all the interfaces had been tested between multiple modules, the whole set of software was tested to establish that all modules could work together correctly as an application or system or package.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Maintenance is necessary to eliminate errors in the system during its working life and to tune the system to any variations in its working environments. Maintenance should meet the scope of any future enhancement, future functionality and any other added functional features to cope up with the latest future needs. The review of the system was done to know fully the capabilities of the system, knowing the required changes or the additional requirements and studying the performance of the system. If a major change to a system is needed, a new project may have to be set up to carry out the change.

Assessment Issues. To address the third research question, the most used evaluation measure in information extraction are accuracy, recall and precision. Recall and precision can be calculated from the confusion matrix according to Manning, C.D (2010):

Recall indicates how many of the items that should be found, are effectively extracted. Precision indicates how many of the extracted items are correct. This is a harmonic mean which in its most general form provides a weight to determine the bias towards recall or precision. More commonly it is used with equal weighting for accuracy, precision and recall.

Effectiveness and Usability Issues. Over and above the assessment issues, user satisfaction with regards to the functionality of the system must also be evaluated in order to address the fourth research question. For this purpose, an evaluation instrument had been formulated to evaluate the web application by using the quality characteristics of the ISO 9126 known as Software Quality Model.

Tools for Data Analysis. Data gathered from different stakeholders had been analyzed carefully that gave the research strong foundation supporting his paper. The following tools were employed.

Site Map. A site map is a list of pages of web site accessible to user. It can be either document in any form used as planning tool for web design or a web page that lists the pages on a website, typically in and hierarchical fashion. Site maps helped the researcher organized its web modules and can easily create navigational links and/or menus.

Frequency Count. This is the most straight-forward approach to working with quantitative data. This is a simple way that the researcher used in counting the number of frequencies for the survey to be conducted.

Five-Point Likert Scale. This method had been used for rating the system. The mean formula had been used to evaluate overall performance. Below is the descriptive interpretation of the rating scale in Table1.

TABLE 1
Numerical Scale and Qualitative Interpretation

RATING	INTERPRETATION
4.51 – 5.00	Strongly Agree
3.51-4.50	Agree
2.51-3.50	Neutral
1.51-2.50	Disagree
1.00-1.50	Strongly Disagree

The comments, suggestions, and recommendations will be noted and may be recommended if found to be valuable in the enhancement of the system.

IV. RESULTS AND FINDINGS

This chapter presents the prototype software implementation of web application, including the results and findings of the study, as well as the respondents' interpretations of those findings in relation to the research questions. To review, the research questions discussed in Chapter 3 are the following:

1. How to develop domain's ontology of terms and file extraction from any type of file to retrieve information.
2. How will the system classify the documents according to the area of exhibit based on the HEI accreditation and assessment requirements.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

3. How to design the dashboard that automatically monitors the accreditation related documents.
4. How the respondents may perceive the level of acceptability the developed web application with respect to chosen criterion under ISO/IEC 9126.

Answering the first research question the researcher studied what was the most appropriate file extractor that would be used in extracting information on a certain document or pdf including the image files. At first, the researcher researched sample applications that would read an image file and just output text extracted from the image. After researching on different extractors, the researcher decided to use the Microsoft Office Document Imaging and Gembox libraries.

Optical Character Recognition (OCR). OCR is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo or from subtitle text superimposed on an image. It is widely used as a form of information entry from printed paper data records, whether passport documents, invoices, bank statements, computerized receipts, business cards, mails, printouts of static-data or any suitable documentation. It is a common method of digitizing printed texts so that they can be electronically edited, searched, stored, displayed, and used in machine processes such as cognitive computing, machine translation, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision.
Microsoft Office Document Imaging is an OCR API that read an image and convert it into text. The Microsoft Office document Imaging tool lets you either open a TIF or TIFF image of a document or scan a document into the Document Imaging tool and then perform Optical Character Recognition (OCR) on the text. A TIF or TIFF image of a document literally means Tagged Image File Format [TIFF] which is then shortened to "TIF." The Microsoft Office Document Imaging tool is a free tool that lets someone with a visual or print disability scan pages into the Document Imaging tool and then make the text readable with their adaptive technology such as a screen reader or TTS/Text. Gembox is used to extract the other types of documents such as Word, Excel, Powerpoint and Pdf as well as hmtl and text files.

Domain Ontology of Term. The researcher developed the ontology by writing down all the related terms based on each area using the old exhibits. The keyword maintenance from the system is being used to store all the terms in the database. The database contains information on how a term will be classified according to its area. The domain ontology consists of eight (8) areas: Community Involvement, Faculty, Curriculum and Instruction, Library, Laboratories, Physical Plants, Student Services, and Administration. For every area, there will be sub-areas to consider building the ontology of terms. Some of area and sub-areas as follows:

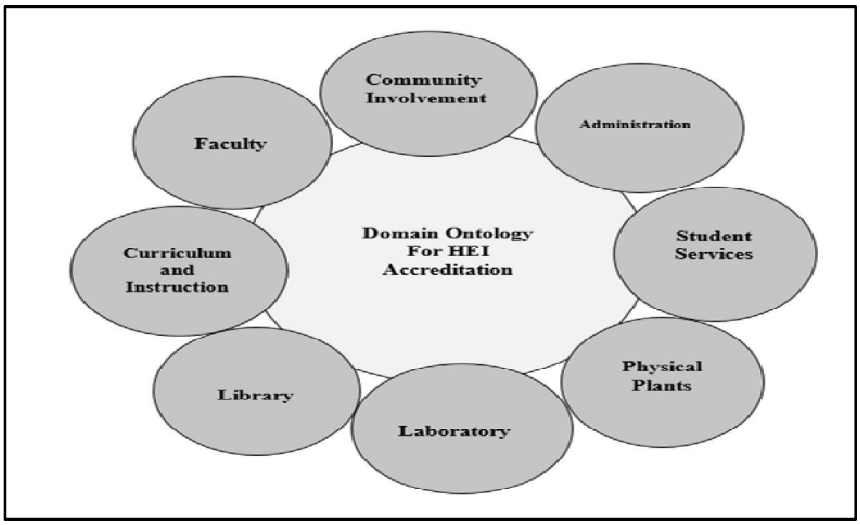

Fig. 3 Domain Ontology for Higher Education Institution Accreditation consists of different areas which will be used in the entire system to classify documents according to its proper area and sub-areas. The implementation of the domain ontology is using database system using SQL server.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Defining Areas of Domain Ontology

There are 8 subject areas that have been defined as the root classes of ontology

1. Community Involvement – contains terms that includes outreach activities and projects
2. Faculty- contains terms that are related to all faculty related activities in the school.
3. Curriculum and Instruction – contains terms related to students' subject area to study.
4. Library – contains terms regarding library activities and procedures
5. Laboratory - contains terms related to laboratory experiments, procedures, budget and purchase.
6. Physical Plants - contains terms related to the physical structure of the HEI.
7. Student Services – contains related terms describing the needed information
8. Administration - contains related term related to development plan, financial statement and annual reports.

In categorizing the extracted data, the researcher tested the accuracy of the application by manually testing several keywords whether the system would pass if the expected words are matched with the actual words.
After the keywords had been cleansed, keyword matching had been implemented for the system to determine which category will the file will be categorized.
The dashboard is connected using SQL server as the backend of the system. It shows the number of documents per Area which will be helpful for the users to monitor the documents uploaded in the system.

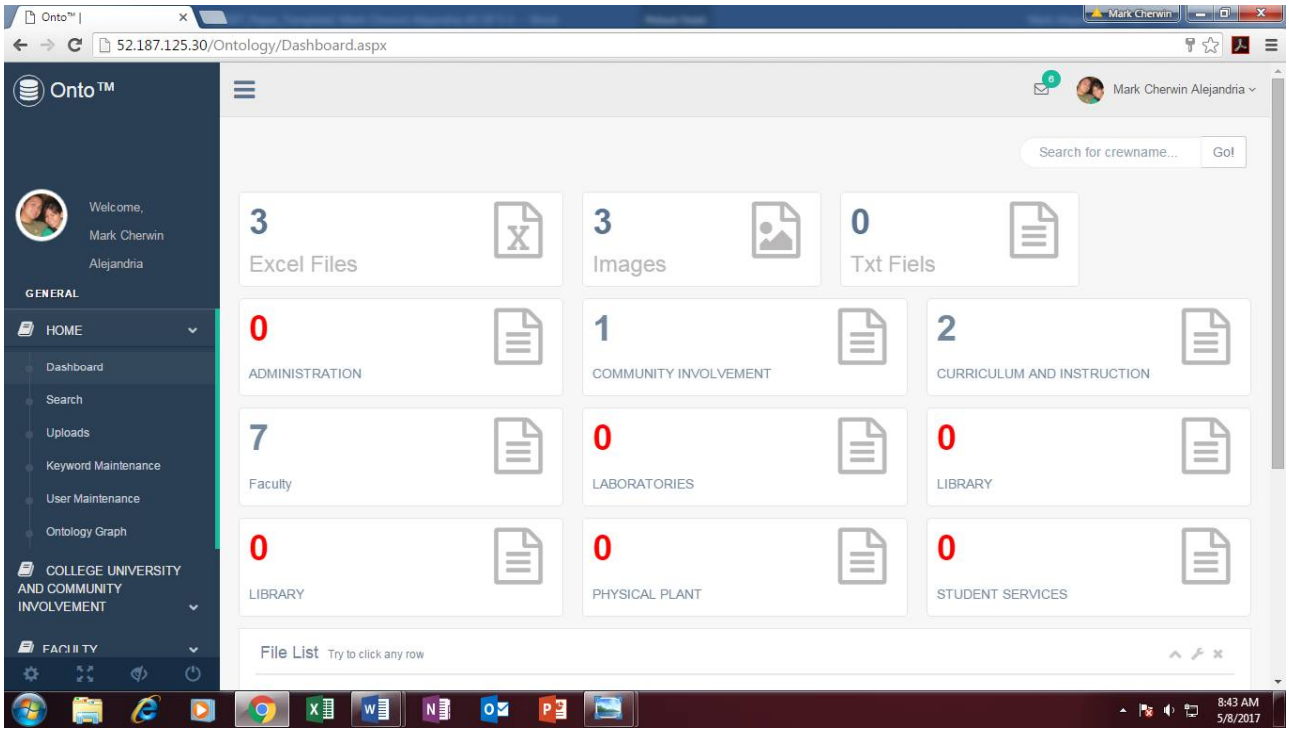

Fig. 5 Screen Shot of the Accreditation Dashboard shows the number of documents per Area to monitor the accreditation document exhibits. It has an alert in red colour informing the users about the lacking documents.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

This acceptability test was conducted with the five (5) respondents which were the members of the accreditation committee that clearly emphasized that the system has all the features, functionality, and modules that satisfied the overall requirements of the respondents.

TABLE 2
Respondent's Assessment on the Prototype Software

Software Characteristics	Mean	Interpretation
Functionality	4.85	Strongly Agree
Reliability	4.70	Strongly Agree
Usability	4.80	Strongly Agree
Maintainability	4.60	Strongly Agree
Portability	4.80	Strongly Agree
Over-All Mean	**4.75**	**Strongly Agree**

Table 4.1 summarizes the evaluation of the respondent's description of the prototype. The overall weighted men is 4.75 which means that the respondents strongly agreed on the different criteria cited above.

V. CONCLUSION & RECOMMENDATION

Individuals tend to find ways how to make life and work less demanding. But making tasks easier needed a lot of sacrifices and dedication. In this study, diverse issues were considered. Surprisingly, people working with accreditations are not mindful that these issues even exist. It is the truth and it ought to be managed and with enough precaution and consideration to circumstance, individuals and the process itself.

The "Semantic Analysis using Ontology for document analysis" web application was developed to help curating accreditation activities simpler. It serves as a reinforcement of past accreditation and a stepping stone and foundation for the next institution that attempted to make a program accredited. It actualized components efficiently intended to drawbacks during the process and made it easier for the user to take advantage of.

The researcher conducted sample applications that read an image file and output text extracted from the image. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. It is a common method of digitizing printed texts so that they can be electronically edited, searched, stored, displayed, and used in machine processes such as cognitive computing, machine translation, key data and text mining.

A. Conclusions

After researching about different extractors, the researcher decided to use the Microsoft Office Document Imaging for images and Gembox for other types. MODI is an OCR API that reads an image and converts it into text. The MODI can be used by installing it in an IDE and add the source packages and libraries into the projects and by writing the code creating for performing the OCR.

After cleansing the extracted data, Keyword density algorithm are used to determine which keyword has the most density in a document as a base to determine the main idea of such document or image.

In categorizing the extracted data, the researcher tested the accuracy of the prototype by manually testing several keywords whether the system will pass if the expected words are matched with the actual words.

Dashboard is also be a useful tool for the users in identifying the needed documents for exhibits.

With an overall mean of 4.75, the acceptability test conducted with the respondents which are the working committee members of the accreditation clearly emphasized that the developed system has all the components, functionality and modules that fulfill the requirements needed by the Accreditation Committee.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

B. Recommendations

The Semantic Analysis of Accreditation Document Using Ontology study has plenty of spaces for further improvements to increase efficiency that future researchers might want to follow through:

1. That the system may be also used by other departments of the College of Arts and Sciences.
2. To use ontology language/application to represent the relationship of each word with a graphical representation
3. Utilization of a centralized repository with document security.
4. Improvement of the quality of the dataset or the keywords used for document classification for more accurate results.
5. Providing decision support based on the recommendations from the Accreditation Agency.

REFERENCES

[1]Adorni, G. et al., "An Ontology-based Archive for Historical Research", University di Genova, 2010
[2] Alicante, A. et al.,"A Distributed Information Extraction System Integrating Ontological Knowledgeand Probabilistic Classifiers", University di Napoli Federico II, 2014
[3] Armour, F. et al.,"Big-Data: Issues andChallenges Moving Forward", In Proceeding of 46th Hawaii International Conference on System Sciences, 2013
[4] Chaudhari, S. and Lade, S. "Classification of News and Research Articles Using Text Pattern Mining", RKDF IST, 2015
[5] Deokar, A. et al., "An Ontology-basedInformation Extraction (OBIE) Framework for Analyzing Initial Public Offering (IPO)", Prospectus, Dakota State University, 2014
[6] Dou, D. and Wimalasuriya, D.C., "Ontology- based Information Extraction: An Introduction and a Survey of Current Approaches", Journal of Information Science, 2010
[7] Dou D. and Wimalasuriya D.C., "Ontology-based information extraction: An introduction and a survey of current approaches", Journal of Information Science, 2010
[8]Elsevier, "International Comparative Performance of the UK Research- Base", Department of Business, Innovation and Skills, 2013
[9] Embley, D. et al, "Ontology-based Extractionand Structuring of Information from Data-Rich Unstructured Documents", School of Accountancy and Information Systems, 2011
[10] Falbo, R. et al.,"ODE: Ontology-based SoftwareDevelopment Environment", Federal University of Espirito Santo Fernando Ferrari Avenue, 2010
[11] Fudholi, D. et al.,"Ontology-based Information Extraction for Knowledge and Validation", La Trobe University, 2016
[13] Han, D. and Stoffel, K.. "Ontology-based Qualitative Case Studies for Sustainability Research", University of Neuchatel, 2014
[14] Hoonlor, A. , "Sequential Patterns and Temporal Patterns for Text Mining", Rensselaer Polytechnic Institute, 2011
[15] Jurafski, D. and Martin J.H., "Speech and Language Processing: An Introduction to Natural Language Processing", Upper Saddle river, 2010
[16] Khoury, R., "Hierarchal Classification in Text Mining for Sentiment Analysis", Lakehead University, 2014
[17] Konys, A., "An Approach for Ontology-basedInformation Extraction System Selection and Evaluation", West Pomeranian University of Technology, 2015
[18] Konys, A., "Knowledge-based Approach to Question Answering System Selection", 7th International Conference on Computational Collective Intelligence Technologies and Applications, 2015
[19] Konys, A., "A Tool Supporting Mining Based Approach Selection to Automatic Ontology", Construction, 9th European Conference on Data Mining, 2015
[20] Lade, S. and Chaudhari, S., "Classification of News and Research Articles Using Text Pattern Mining", RKDF IST, 2015
[21] Lewis, D.D. , "An Evaluation of Phrasal andClustered Representations on a Text Categorization Task" , Proc. 15th Ann. Int"l ACM, 2015
[22] Manning, C.D. et al., "An Introduction toInformation Retrieval", Cambridge University, 2010
[23] Martin, M.,"Ontology-based Data Integrationbetween Clinical and Research Systems", Institute forMedical Informatics, 2015
[24] Martin J.H. and Jurafski, D., "Speech andLanguage Processing: An Introduction to NaturalLanguage Processing", Upper Saddle river, NewJersey, 2010
[25] Pardede, E. et al., "CODE (Common OntologyDevelopment): A Knowledge Integration Approachfrom Multiple Ontologies", In Proc. IEEE 28th AINA, 2014
[26] Rahayu, W. et al.,"A Data-driven Dynamic Ontology", La Trobe University, 2015
[27] Sebastiani, F. "Machine Learning in Automated Text Categorization", ACM Computing Surveys, 2012
[28]Toledo, C. et al.," An Ontology-drivenDocument Retrieval Strategy for OrganizationalKnowledge Management Systems", Electronic Notes in Theoretical Computer Science, 2013
[29] Weikum, G. et al., "Combining Information Extraction and Human Computing for Crowdsourced Knowledge Acquisition", Max Planck Institute for Informatics, 2014
[30] Witten, I. et al., Domain-Specific Key Phrase Extraction, University of Waikato, 2010
[31] Wimalasuriya, D.C. and Dou, D., "Ontology-based Information Extraction: An Introduction and aSurvey of Current Approaches", Journal of Information Science, 2010
[32] Wimalasuriya D.C. and Dou D., "Componentsfor Information Extraction: Ontology-BasedInformation Extractors and Generic Platforms, CIKM"10, 2010
[33] Zhang, J. and El-Gohary, N.,"Automated Regulatory Information Extraction from Building Codes Leveraging Syntactic and Semantic Information", Construction Research Congress, 2012