Islamic Resources Big Data mining, Extraction and Archiving

Mohamed Menacer[1], Abderahmane Menacer[2], Amar Arbaoui[3]
[1, 3]Noor Research Centre, Taibah University, College Computer Science & Engineering, Madinah, Saudi Arabia
[2]Saad Dahlab University, Institute d'Electronique, Blida, Algeria

Abstract: Big data technologies is an evolving term that describes the variety of techniques for searching, collecting, processing, analysing, and storing a large amount of structured, semi-structured and unstructured data that can be mined for information. This is certainly the case for Islamic resources and materials that are widely available as structured, mainly unstructured and unrelated information on the web, this can be referred to as Islamic big data. The aim of this paper is to present the work that has been carried out to search, extract and collect related Islamicinfo on a big data scale that are relevant to our study, after due copyright issues are dealt with and permissions are sought, if any. The findings would be madeavailalble as structured and related info which is of benefit to the relevant audience in a form of knowledgebase, searchable resources, as well as a source for analytical, qualitative, and quantitative analysis of Islamic web resources in general. Semi-automatic search and auto-extraction techniques that have been used in this study will be presented as well as preliminary results. This work is funded by NOOR Research Centre, Madinah-Saudi Arabia, where its main focus is on IT development for Quran and Islamic Sciences.

Keywords: Big data, Islamic big data, data mining, search, collection, Content extraction, Quran resources.

I. INTRODUCTION

The rise in information technology has led to an abundant amount of structured and unstructured data on the web, which requires more sophisticated data processing and storage systems. The increasing volume and detail of information captured by individuals and organizations, the rise of multimedia, social media, and the Internet of Things will have exponential growth in data, for the foreseeable future, that is of great value to business, science, government, and society in general. Big Data has created the need to develop novel techniques to mine, analyse, organize and save large and complex data sets and information, otherwise difficult to process and manage with conventional tools or traditional data processing techniques. This will become a key basis for new waves of productivity growth, competition, innovation, and consumer surplus. Furthermore, big data sophisticated analytics can substantially improve decision-making, that can be used to improve the development of the next generation of products and services. The societal benefits of these techniques and services are immeasurable, in transforming how people search and make use of information on a daily basis [1, 2, 3].This is certainly the case for the huge large amount of Quran and Islamic resources, materials and information that are widely available over the internet, however lagging behind in terms of a structured information, digital content and resources [4,5, 6].

This paper aims to present the tools, processes and procedures for the search and collection of Quran and Islamic data (books, research, papers, websites, articles, conferences, journals, websites, …) that is available over the web that would allow for the creation, manipulation and management of large data sets in an organized manner in a form of a structured knowledgebase or/and easily searchable information, with ever evolving dat, content and resources [7]. The above work is being implemented under a research project, which is at its initial stages.

II. ISLAMIC BIG DATA

Quran and Islamic digital information over the web has noticed a huge increase due to the technological advancements in Information technologies in the areas of publishing, indexing, searching and multimedia. There arevariety ofIslamic resources (Ancient Islamic manuscripts, books related to Quran, research papers, magazines, journal papers, websites, …etc) that are available in many places and in different languageson the internet. Furthermore, The vast and scattered Islamic resources are available in different formats from different sources makes it difficult to consistently and comprehensively mine and collect such information [8,9].

In this work, a strategy for search and data collection has been developed in order to gather, structure and organize the selected information that is available over the web. Due to the complexity of the task, limited and specific information and resources have been selected for search and acquisition such as books, articles, journals, conferences, and websites; in two different languages: Arabic and English. However, due to the sensitive issueof the information and resources being collected, exhaustive inspection for validation and authentication has to take place. This is a crutial part of the process that would make this work reliable with trusted information.

III. DATAMINING& ACQUISITON METHODOLOGY

A consistent search methodology has been developed in order to collect and gather the most relevant, authentic and trusted materials and resources on Quran and Islamic sciences. At the initial stage of this work and in order to identify the types of resources available for collection, information was manually gathered from different Islamic websites, and other resources such as libraries, universities and organizations. Some of the problems encountered during this process are: the information collected is incomplete, unorganized, and in different formats and languages. Due to these problems, processes for structuring and organizing such material were defined in order to provide proper guidelines for data mining and collection processes. The main phases of the data collection structure are shown in Figure 1.

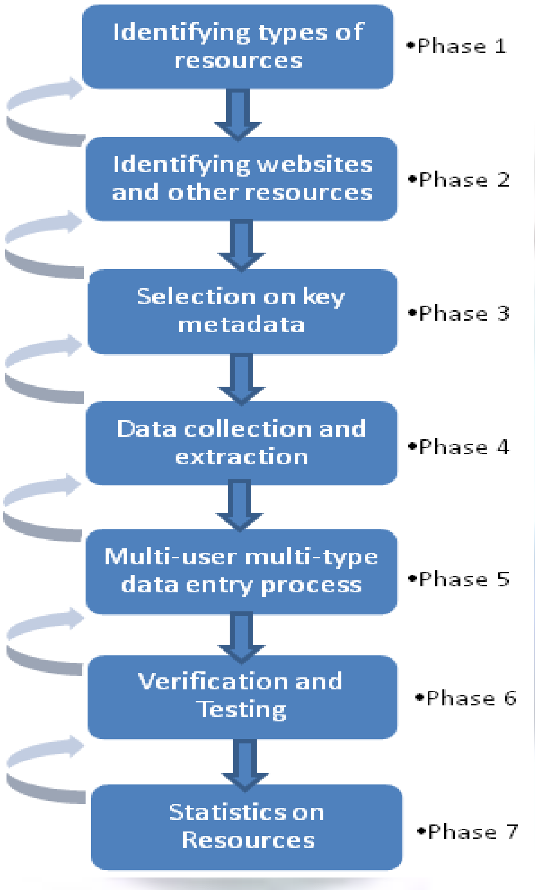

Identifying types of resources
•Phase 1
Identifying websites and other resources
•Phase 2
Selection on key metadata
•Phase 3
Data collection and extraction
•Phase 4
Multi-user multi-type data entry process
•Phase 5
Verification and Testing
•Phase 6
Statistics on Resources
•Phase 7

Figure 1: Phases of Data Collection Structure

The structure, in Figure 1, emphasizes the need for checking and verification between all the phases of the data collection. The data collection and entry processes are performed by data entry officers and verified by a control officer who checks the information entered:

1. **Identifying types of resources:** The initial phase in the data collection structure process is to identify the types of resources which are relevant and authenticated about Quran and Islamic related info and resources.

2. **Identifying lists of resources:** From extensive search on the internet, thousands of Islamic related websites were scrutinized in order to create lists for the different relevant resources. In addition, Some other resources were found from visits to some public and university libraries, in Saudi Arabia, that provided some informationon books and journals not readily available online. Furthermore, specific and relevant info and content within the same website are selected during this phase. For the latter, emphisis is made on the copyright of the data/content being targetted, in case it is not clear permissions are sought before any data extraction is made. This process has been very tedious and time consuming due to the manual aspect being involved.

3. **Selection on key metadata:** In an attempt to construct the required metadata fields for relevant searches, metadata lists associated with each type of resources were created. The purpose of selecting key metadata is to ease and efficiently search for appropriate and relevant Quran and Islamic resources. This is a key aspect to minimize the unwanted info and resources.
4. **Data collection and extraction of resources:** in this phase, the gathered lists of websites of interestand the selected metadata are used for automatic extraction and download of relevant info/data/content. A Dedicated system has been developed for the automatic extraction and filtering of the collected data.

5. **Multi-user multi-type data entry process:** A multi-user with multi-type data entry form has been designed for logging the various websites with their metadata inot the automatic extraction system. This process has mainly been done manually. However, an automated process is being developed for this phase forthe resources where information may be automatically extracted.

6. **Verification/Testing:** The verification and testing phase is an ongoing process and is performed after each phase by the data entry officers and then verified and tested by the control officer(s), who is in charge of authenticating the accuracy of the data entry process. The data collection system has been developed to easily trace back any entries or modifications.

7. **Statistics on resources:** The data collection system provides the administrator (control officer) with up to date statistics on all info and records being entered as well as data entry officer, types of resources, resources records, time, date, etc. This statistics user interface will help in tracking resources and errors which may occur during the data entry and collection processes.

IV. AUTOMATIC EXTRACTION OVERALL DESIGN AND DEVELOPMENT

A data mining and collection system has been developed based on number of tools for the search, collection, automatic extractions, and filteringof the various types ofdata and resources. Figure 2, shows the design process for data collection, extraction, filtering and archiving of relevant resources and materials. Furthermore, the design allows for manual and automatedprocedures to efficiently organize all collected materials and resources in order to simplify and enhance the acquisition, classification, and archiving processes. The system consists of a structured database with data acquisition entry forms, automatic extraction, filtering and verification tools.

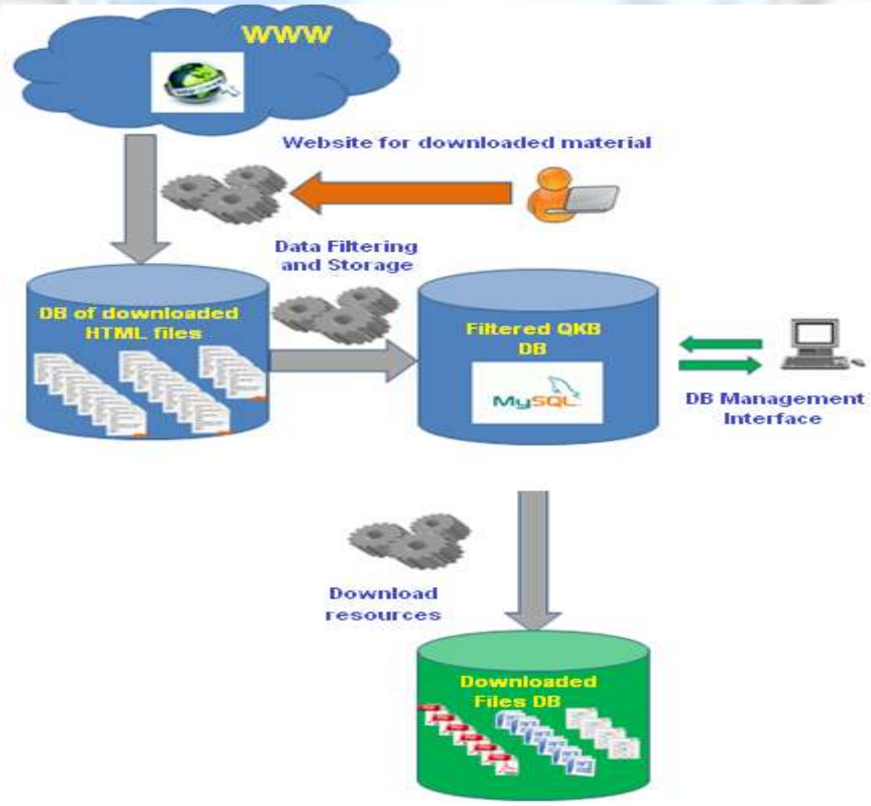

WWW
Website for downloaded material
Data Filtering and Storage
DB of downloaded HTML files
Filtered QKB DB
MySQL
DB Management Interface
Download resources
Downloaded Files DB

Figure 2: The design process for data collection, extraction and archiving

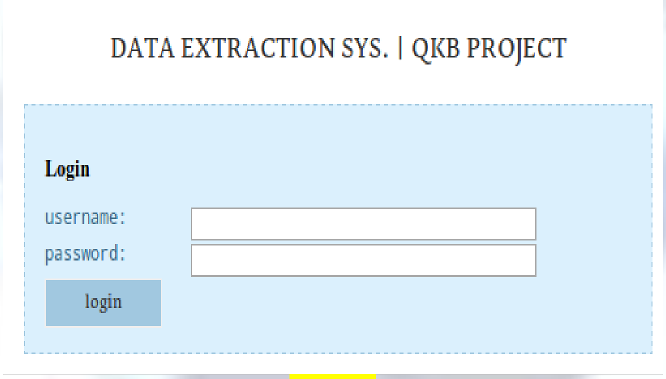
V. Implementation and Discussion

The implementation phase of this work has been done in several stage:

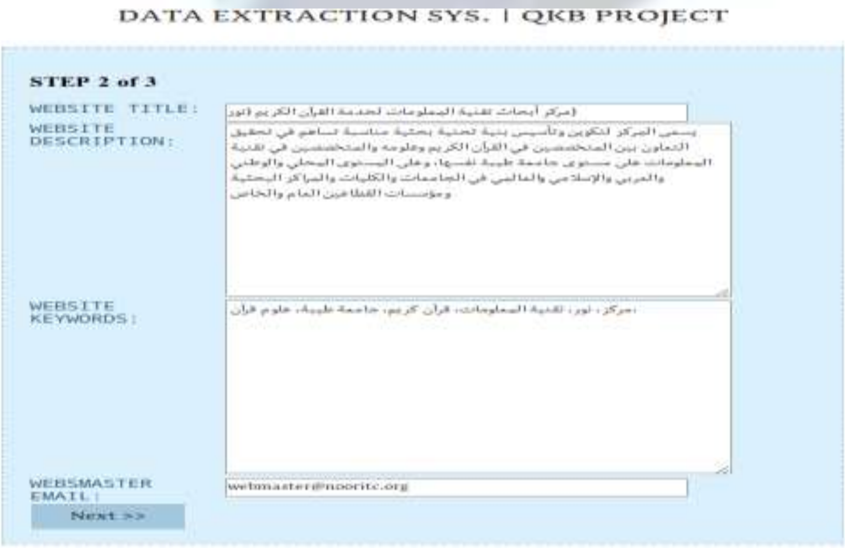
- Gathering of relevant info and metadata for data enty, which is fully manual, at this stage of the work, that is time consuming and requires a number of human resources.
- Automatic extraction process that required several servers to be run at the same time with close monitoring. It also required a large internet bandwith and took months for execution.
- Filtering and cleaning process for unwanted and uncorrect data. This has been a key stage in fine tuning the automatic extraction system, as well collecting meaningfull data.
- Validation and authentication process, which is an going process, ensures the reliability, classification and authentication of the data being collected. Furthermore, irrelevant but useful data is as well indexed and classified in order to make it useable and functional. The later issues needs to be improved and automated in order to optimize and its efficient use on a Big data scale.

Figure 3 shows the different steps for the implementation and execution of the automatic extraction system:

- **Step 1:**Each user, from data entry to process and data monitoring, is logged in in order to provide a high reliability, consistency and monitoring for the whole data collection process.
- **Step 2:** Data entry of the relevant info, metadata and URL link of all selected sources for data collection. Manual entry is necessary at the first stage of the project, however, and algoritm is being developed for and automated website resources search at later stages
- **Step 3:** Specific task(s) are defined for the data extraction from a specific website. These tasks would be the targeted data and content to be extracted, that can be monitoring during and after extraction.
- **Step 4:**a code editor is displayed, before execution, with the details of a specific task for reviewing and amending the codes, if necessary, by a specialized individual or developer. This is due to the fact that different forms and formats can be found with different web technologies on one hand, to speed up and ease the code checking process on the other hand.



STEP 1



STEP 2 of 3

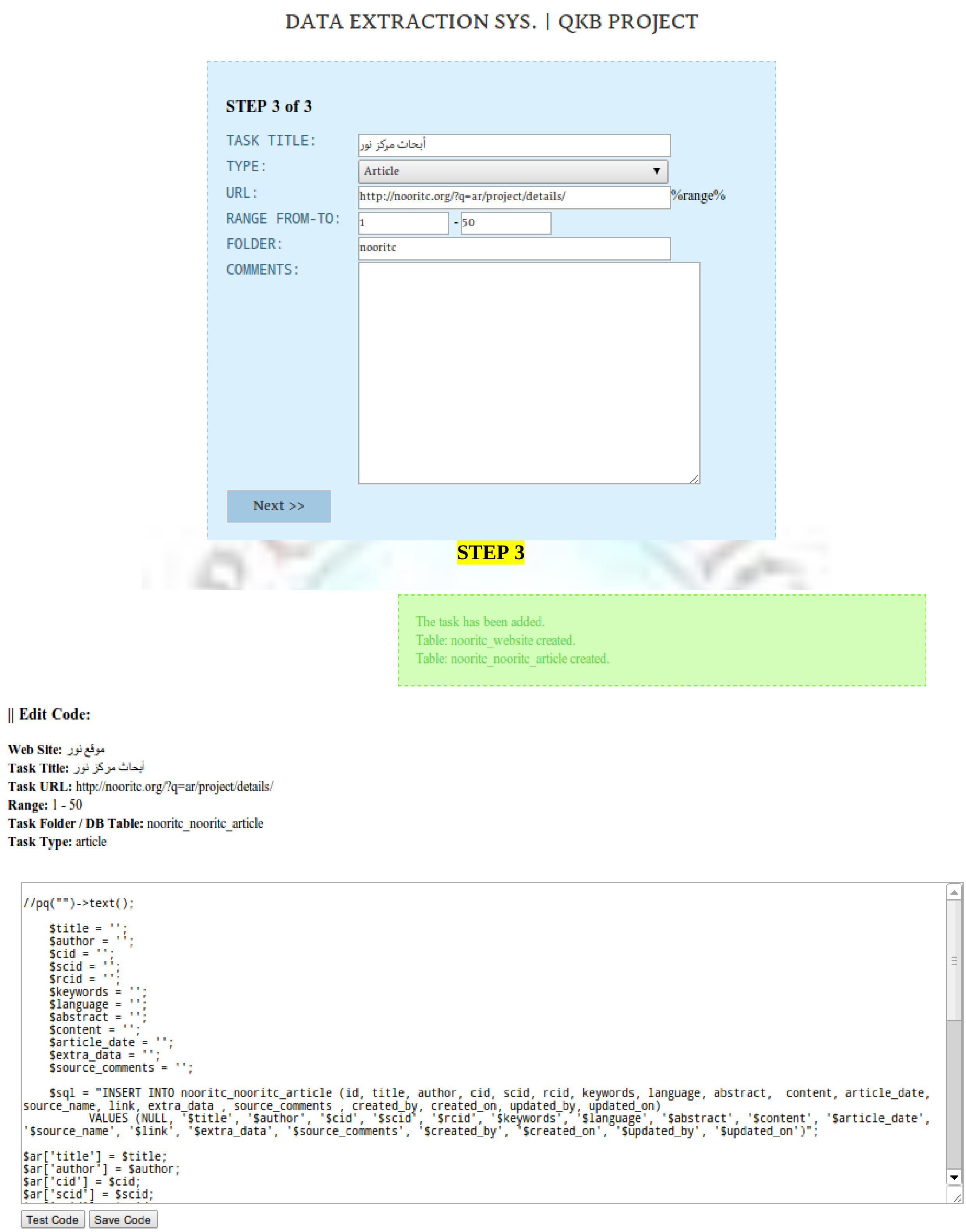
WEBSITE TITLE: (مركز أبحاث تقنية المعلومات لخدمة القرآن الكريم (نور

WEBSITE DESCRIPTION: يسعى المركز لتكوين وتأسيس بنية تحتية بحثية مناسبة تساهم في تحقيق
التعاون بين المتخصصين في القرآن الكريم وعلومه والمتخصصين في تقنية
المعلومات على مستوى جامعة طيبة نفسها، وعلى المستوى المحلي والوطني
والعربي والإسلامي والعالمي في الجامعات والكليات والمراكز البحثية
ومؤسسات القطاعين العام والخاص.

WEBSITE KEYWORDS: مركز، نور، تقنية المعلومات، قرآن كريم، جامعة طيبة، علوم قرآن

WEBSMASTER EMAIL: webmaster@nooritc.org

STEP 2



DATA EXTRACTION SYS. | QKB PROJECT

STEP 3 of 3

TASK TITLE: أبحاث مركز نور
TYPE: Article
URL: http://nooritc.org/?q=ar/project/details/ %range%
RANGE FROM-TO: 1 - 50
FOLDER: nooritc
COMMENTS:

STEP 3

The task has been added.
Table: nooritc_website created.
Table: nooritc_nooritc_article created.

|| Edit Code:

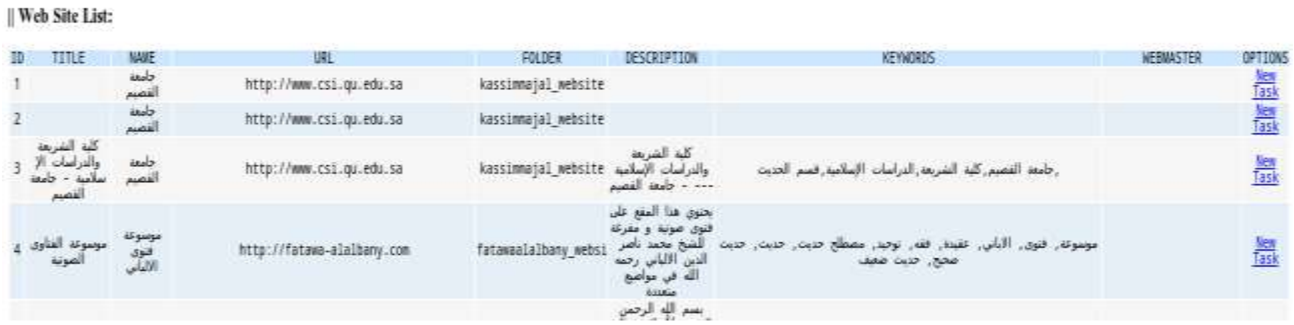
Web Site: موقع نور
Task Title: أبحاث مركز نور
Task URL: http://nooritc.org/?q=ar/project/details/
Range: 1 - 50
Task Folder / DB Table: nooritc_nooritc_article
Task Type: article

```
//pq("")->text();

    $title = '';
    $author = '';
    $cid = '';
    $scid = '';
    $rcid = '';
    $keywords = '';
    $language = '';
    $abstract = '';
    $content = '';
    $article_date = '';
    $extra_data = '';
    $source_comments = '';

    $sql = "INSERT INTO nooritc_nooritc_article (id, title, author, cid, scid, rcid, keywords, language, abstract,  content, article_date,
source_name, link, extra_data , source_comments , created_by, created_on, updated_by, updated_on)
          VALUES (NULL, '$title', '$author', '$cid', '$scid', '$rcid', '$keywords', '$language', '$abstract', '$content', '$article_date',
'$source_name', '$link', '$extra_data', '$source_comments', '$created_by', '$created_on', '$updated_by', '$updated_on')";

$ar['title'] = $title;
$ar['author'] = $author;
$ar['cid'] = $cid;
$ar['scid'] = $scid;
```

STEP 4

Figure 3: Steps for the automatic extraction process of relevant resources from trusted websites.

Figure 4, presents some of the results from the monitoring panel for extracted data and content from selected websites after performing the automatic extraction process. The info being displayed shows the different details of the data extracted as well as the progress of the extraction tasks, which is very useful for monitoring purposes, analytical analysis, as well as for future updates and additional tasks that can be performed automatically at later dates.



|| Web Site List:

ID	TITLE	NAME	URL	FOLDER	DESCRIPTION	KEYWORDS	WEBMASTER	OPTIONS
1		جامعة القصيم	http://www.csi.qu.edu.sa	kassimmajal_website				New Task
2		جامعة القصيم	http://www.csi.qu.edu.sa	kassimmajal_website				New Task
3	كلية الشريعة والدراسات الإ سلامية - جامعة القصيم	جامعة القصيم	http://www.csi.qu.edu.sa	kassimmajal_website	كلية الشريعة والدراسات الإسلامية --- - جامعة القصيم	جامعة القصيم,كلية الشريعة,الدراسات الإسلامية,قسم الحديث,		New Task
4	موسوعة الفتاوى الصوتية	موسوعة فتوى الالباني	http://fatawa-alalbany.com	fatawaalalbany_websi	يحتوي هذا المقع على فتوى صوتية و مفرغة للشيخ محمد ناصر الدين الالباني رحمه الله في مواضيع متعددة	موسوعة, فتوى, الالباني, عقيدة, فقه, توحيد, مصطلح حديث, حديث, حديث, حديث صحيح, حديث ضعيف		New Task
					بسم الله الرحمن			

Figure 4: Monitoring Panel and results of extracted data and content after performing the automatic extraction process.

VI. Conclusion

Although the research project is still under development at its initial stages, overall concepts of the automatic extraction system has been presented. A clear methodology has been put in place for collecting and gathering the most relevant, authentic, and trusted data, materials and resources on Quran and Islamic sciences on a large scale over the web. An automatic extraction system for collection has been developed based on number of tools for the search, acquisition, and filteringof the various types ofdata and resources.Although, most efforts have been made to collect relevant data, there were a large number of unwanted/irrelevant data that have been collected. The latter materials are being filtered out, sometime manually, which is time consuming and requires more human resources or/and further dedicated data processing techniques. The analytical and statical on the data collected are being processed and preliminary results show an overall satisfactory for this first phase of the research work. Furthermore, this part of the project is still in progress with the final phases to be comprehensively implemented.Finally, this work is a preliminary attempt to collect, organize, and archive Islamic and Quran related data and resources on a Big Data environmentr. As the work progresses it is expected to encounter new and complex challenges that would make this research program even more ecxiting, and discover the abandunce and richness in the structured and organized manner of thehuge amount of data and resources being collected that be would of great benefit to individuals and organitions alike.

Acknowledgment

The Authors would like to acknowledge and thank NOOR Research Centre for their funding, and unconditional support of this research project (Project ID: NRC1-167).

References

[1]. R. Akerkar , “Big Data computing”, First Edition, Chapman and Hall/CRC, 2013.
[2]. “Big data: The next frontier for competition”, http://www.mckinsey.com/features/big_data, last viewed 5 September, 2014.
[3]. “Big data”, http://en.wikipedia.org/wiki/Big_data, last viewed September, 2014.
[4]. G. Larsson, T. Hoffman, “The impact of the Internet on the Qurʾān,”, Muslims and the New Information and Communication Technologies, Springer Verlag, 2012
[5]. “How Digitization Has Changed the Cataloging of Islamic Books”, https://researchblogs.cul.columbia.edu/islamicbooks/, last viewed May 2014.
[6]. “How ‘big data’ is changing our lives” http://islam-science.net/how-big-data-is-changing-our-lives-637/, last viewed September, 2014.
[7]. Amar Arbaoui, Yasser M. Alginahi, Mohamed, " Menacer, Strategies for Collecting Electronic Resources on the Qur'anic Researches", International Journal on Quranic Research, ISSN 2180-4893, Volume 3 No 1 – 2013.
[8]. M. Menacer, A. Arbaoui, " Content extraction of Quran Interpretation (Tafseer) Books for digital content creation and distribution ", ICMMP’2013 International Conference on Multimedia processing, Sousse, Tunisia, 22-24 June 2013.
[9]. Y. Alginahi, M. Menacer, A. Arbaoui, "Methodology and Processes for Collecting, indexing, Structuring Material and Resources of Quran and Islamic Studies" , WSCAR’2014World Symposium on Computer Applications and Research, Sousse, Tunisia, June 2014.