Towards Establishing an ETD for Egyptian Universities

M. F. Tolba T. M. Nazmy Y. M. Abd El-Latif E. Abdelwahab

Ain Shams University Faculty of Computer & Information Sciences
Abbassia, Cairo, Egypt

Abstract — **Many universities and libraries throughout the world are now making digitized versions of traditional (print) dissertations available online. The Ain Shams University Information Network, ASUNET, is the repository of the scientific theses published by the Egyptian universities since 1992 with more than one hundred fifty thousand, bilingual, Arabic and English documents and more than ten thousand new deliveries per year. The transformation of this huge theses library into an electronic theses and dissertation database (ETD) is an important and big challenge. This paper gives a survey on similar theses libraries world-wide, modern search and repository technologies, problems encountered, and accessibility issues. This is followed by describing the current situation of the ASUNET effort and ways to cooperate with other local projects in order to achive best results. A proposal for the construction of a modern theses library is finally presented.**

Index Terms — **ETD, meta-data, catalog, indexing, subject vocabularies, ontology, conceptualization, ASUNET Library, Science Grid, Union Databases.**

I. INTRODUCTION

Putting theses and dissertations online offers more than just the possibility to manage, store, organize and disseminate digital materials created by an institution and its community members [1-5]. By developing new services which take advantage of the characteristics of electronic media new values can be added to the original documents at the same time as some shortcomings can be overcome [1]. It is obvious that such en endeavor will be most welcome in the environment of Egyptian universities which is still lacking many modern infrastructure and logistic means.

In developing a strategy for the development of an ETDs in Egypt we do wish to start with metadata. Using submission software, students would be able to create their own metadata, which is quality-controlled by the Library. Also in the loop, inevitably, is the authority responsible for validating the approved thesis. Interaction between students and supervisors goes on throughout the course of the degree program. Finally, the system outputs are the metadata, formatted as required for various agencies, and for the ETD itself.

Our expectation is that an XML schema – or perhaps a number of schemas - will be developed for Egyptian theses and dissertations, possibly based upon schemas which already exist for use in other countries. A schema will describe each thesis according to its various structural elements, and should support the export of metadata in all of the various formats required, while at the same time describing the full text of the thesis. In other words, PDF is not likely to be sufficient in the longer term. Using XML provides us with a non-proprietary format, with greater scope for database storage of deconstructed documents, greater search flexibility, and the possibility of preserving the 'raw' source of the document in addition to various other advantages to be discussed in the course of the paper.

The more challenging task may be to find universities which are willing to allow ETDs to be created in their institutions, and to work with us in a joint project, as pilot sites. Much of our work will be on the political and cultural changes needed in institutions in order to prepare them for the inevitable future context of ETDs. For some institutions, moving to an environment in which the electronic thesis or dissertation is the authoritative copy, the one which is preserved and used, may seem a huge step which is still years away. Even in the US, the numbers of institutions which have made provision for ETDs is still relatively small, though growing almost daily.

The paper is organized as follows. The next section will describe the ETDMS standard which is intended to be the basis upon which we build our system. Section 3, deals briefly with data interchanges standards and formats and information storage. This is followed by a description (Section 4) of digital libraries strategies in the Egyptian repository for theses and dissertation. In Section 5 modern standards of information retrieval are outlined before concluding with our main technical and strategic recommendations.

II. THE ELECTRONIC THESES AND DISSERTATION METADATA STANDARD (ETDMS)

The Electronic Thesis and Dissertation Metadata Standard (ETDMS) was developed in the UK in conjunction with the Networked Digital Library of

Theses and Dissertations (NDLTD), and has been refined over the course of years. The initial goal was to develop a single standard XML DTD for encoding the full text of an ETD. Among other things, an ETD encoded in XML could include rich metadata about the author and work that could easily be extracted for use in union databases and the like. During initial phases it became clear that the methods used by different institutions to prepare and deal with theses and dissertations would make it all but impossible to agree on a single DTD for encoding the full text of an ETD. Many institutions were unwilling or unprepared to use XML to encode ETDs at all [2], [3]. Thus, instead of an XML DTD for encoding the full text of an ETD, ETDMS emerged as a flexible set of guidelines for encoding and sharing very basic metadata regarding ETDs among institutions. Separate work continues in parallel on a suite of DTDs, building on a common framework, for full ETDs.

ETDMS is based on the Dublin Core Element Set ([5],[6]) but includes an additional element specific to metadata regarding theses and dissertations. Despite its name, ETDMS is designed to deal with metadata associated with both paper and electronic theses and dissertations. It also is designed to handle metadata in many languages, including metadata regarding a single work that has been recorded in different languages. The ETDMS standard [4] provides detailed guidelines on mapping information about an ETD to metadata elements. ETDMS already is supported as an output format for the Open Archives interface to the Virginia Tech ETD collection. UKs NDLTD strongly encourages use of ETDMS.

In this context it is important to enlighten the concept of resource sharing between libraries and similar institutions, because it provides the only basis for applying any technological standard like the ETDMS ([7],[8]).

‘Library Resources’ is the term that applies to personnel, material, functions or activities available in a library for satisfying the human needs & demands to acquire their desired knowledge. Library co-operation is a very old concept and a form of resource sharing. The new object of resource sharing has changed the old concept due to multi-dimensional growth of published documents through R&D activities in recent past, cost of the information, advancement of newly invented technologies for information processing and dissemination, etc.

For better utilization of resources, participating libraries should come together and co-operate in two broad areas [7] : (a) developing the collection on shared basis; and (b) improving services for exploiting such collection. The conventional library, and especially in Egypt, is seriously affected by some barriers of information communication, such as indifference of employees, conservative attitudes, distance, language, cost, time, etc. for inter-library loan. And there are also several constraints to resource sharing in the print environment as it existed till today: (a) open access to shared resource is not possible; (b) service depends upon library performance; (c) access to shared resource at a cost; (d) legal issues; (e) non-availability of library financial resources ; and (f) authenticity of collected information resources on the Internet. All this plus a trend to work in "isolated islands" is a very serious managerial challenge to be met by anybody wanting to realize an efficient ETD system.

III. DATA INTERCHANGE STANDARDS AND FORMATS, DATA STORAGE

This section briefly reviews possible ETD document formats in addition to modern storage devices. This is done for the sake of completeness and may be skipped by technically versed readers (see [9],[10]).

A. Formats

Traditional journals were available in one manifestation print. With the increased use of computers and the explosion of the Internet, however, articles are now produced and published in multiple manifestations.

Most articles are available electronically in Adobe Portable Document Format (PDF), PostScript Format (PS), and Standard Generalized Markup Language (SGML) formats.

Adobe's PDF proprietary format, which employs the Acrobat suite of software products to be created, edited, viewed, etc. Is printing device-independent, and supports e-publishing using sophisticated formatting and graphics including embedded links, annotations, thumbnails of pages, and chapter outlines for direct access. Adobe has indicated the intent to incorporate structure as well as layouts into PDF by extending it to encompass SGML. PS is an Adobe-developed programming language of 420 format command operators (Level-2) which control printing (but not screen display) and allow formatted printing on any printer from any platform (Windows, UNIX, etc.). Encapsulated PostScript (".eps") are subroutines included in PostScript files, usually used for images produced with a non-PostScript package.

Compared to page description languages, structured information mark-up languages describe the information (content) and structure, not the layout. They are device and processing platform independent and facilitate automatic indexing by describing headings, chapters, paragraphs, footnotes, etc. An SGML (Standard Generalized Mark-up Language (ISO Standard 8879-1986) document has three elements: the Declaration (describes processing environment needed); the Document Type Definition (DTD) (a defined tag set that forms a template for describing the structure and content of a specific type of document); and the Document stream itself. SGML is independent of any system, device, language or application, and, because it separates document content definition from presentation, it allows information to be accessed or presented in ways not predicted at the time of mark-up. SGML viewing software (e.g. Panorama) parses/interprets the SGML document content according to its DTD instructions. SGML is anticipated to be a key standard in digital library development.

XML is a simple, reduced subset of SGML designed (in 1996) for ease of implementation and interoperability with both full SGML and HTML. It is simpler than SGML (reducing a 500-page reference to 26 pages). Unlike HTML, XML supports (optionally) user-defined tags and attributes, allows nesting within documents to any degree of complexity, and can contain an optional description of its grammar for use by applications that need to perform structural validation. Every valid XML document will be a conformant SGML document. XML is not backward compatible with HTML documents, although those conforming to HTML 3.2 can easily be converted. It is not intended to supplant HTML but to complement it.

HTML is a reduced tag set version of an SGML DTD that provides a set of platform-independent styles (defined by tags) used to define the components of a Web document. HTML 2.0 is an Internet Engineering Task Force (IETF) standard. While HTML tags are primarily structure-related, there are increasingly accepted tags for specifying presentation and layout.

DHTML (dynamic HTML) denotes recent developments by both Netscape and Microsoft that use a combination of Cascading Style Sheets and a scripting language such as Visual Basic script or Javascript to merge the HTML document with the style sheet. It supports greater creative control over the visual presentation of an HTML page and allows the page to respond dynamically, without a call to the server, to user-generated events.

The DSSSL, Document Style, Semantic and Specification Language (ISO 10179), is a standard associated with SGML that specifies the rules for a non-proprietary language to govern the appearance and style for the logical components (e.g. chapter headings) defined by SGML.

CSS is a new approach to increasing control over the visual formatting of HTML documents (e.g. spacing, colors, backgrounds, choice of fonts, drop shadows, layering, relative and absolute positioning, on/off visibility of options, choice of media such as print, display, braille, aural). CSS tags are in a separate document or part of the document (rather than being embedded in the text as with traditional HTML), so can be changed, updated across multiple documents quickly. Cascading style sheets can be cached locally and reused, so their deployment can result in bandwidth/response time gains.

SGML and PDF have advantages and disadvantages. From an archivist's standpoint, SGML presents two key advantages over PDF:

1. Because SGML is human-readable and non-proprietary, there is a higher probability that files will be usable in the time frame of the proposed archive (75 or more years). There is no guarantee that Adobe Acrobat (or other software that can render PDF files) will be available for the computers in use at a distant time in the future.
2. Because SGML retains content structure, it will be possible to index and search the article content, and to construct links to databases that do not exist today.

For these reasons, the focus of most reports is archive of SGML files. However, Inera™ recommends that if space and logistics allow, PDF files has also to be included in the archive for the following reasons:

1. PDF files retain the original visual presentation, preserve a different aspect of the print journal and therefore may be more useful for some research activities.
2. Some Non-Article and Other Content is available only in PDF format. If the archive does not accept PDF, this content will be lost.

B. Storage Media

Storing master files on high quality media is a vital step in any ETD system. This can be done on one of the following types of media:

- *Optical storage*
- *Magnetic storage*
- *Micro formats storage*

In 1980 a standard known as Red Book that defined music compact discs in terms of sampling rate (44,100 samples per second), range of values (65,536), and physical format was developed. In 1983 the Yellow Book standard retained the physical format of Red Book, added more error correction, and allowed two data structures: Mode 1 (ISO 9660), best for data unforgiving of error such as computer programs or databases; and Mode 2, for data more tolerant of error such as audio, video and graphics. Green Book is a standard, based on Yellow Book Mode 2 that defines the disc, its contents, special compression methods for audio and visual data, an interleaving method for audio, video and textual data, and hardware and software system designed to interact with television and stereo systems and, more recently, with the Web. Orange Book, Part II, created in 1988 is a specification that supports one-time recording onto discs of all types of random access data, using a laser-sensitive dye layer in the disc make-up.

Orange Book, Part III, 1994 supported erasable disc recording based on a phase-change- sensitive film in the disc make-up. It was not backward compatible with CD players or CD-ROM drives due to a difference in media reflectivity. Finally, the modern generation of optical disc formats has higher storage capacity (8-15 times more) than its CD equivalents and is called: DVD.

In contrast to optical storage magnetic storage media such as magnetic tapes, diskettes, and cartridges are prolific and largely proprietary. Widely used Micro format storage media include 16 mm, 35mm, 70 mm microfilm and 105 mm microfiche.

IV. THE EGYPTIAN REPOSOITORY FOR THESES AND DISSERTATION

A. Historical Background

The importance of establishing a central Egyptian repository of scientific thesis was recognized as early as 1967 where the "library of post-graduate and research documents" was created at the locations of Ain Shams University. In 1992, when the Ain Shams University Information Network (ASUNET) was created to build a comprehensive microfilm database. In view of the large numbers of available documents, contributing to the biggest contemporary collection of Arabic graduate-studies texts in the region, the aim of this effort was to give researchers and other interested parties a seamless opportunity of accessing an extensive archive of scientific materials dedicated to R&D within the Arab region. With time this repository has increased to reach the proud number of 200000 documents in 2006, increasing by an estimated 15000 in each consequent year. This increase and the necessity to adopt new technologies enabling world-wide beneficial use of this collection encouraged us to the project idea presented here.

The main goals of the proposed project are:

1. Converting ca. 11000 Microfilm containing more than 150000 scientific theses into an electronic readable form.
2. Indexing the thus created electronic archive.
3. Building a modern electronic search functionality which enables retrieving information from abstracts as well as full-text documents.
4. Offering both online and offline search functionalities
5. Integrating the created archive with similar databases world-wide
6. Providing accessibility to this repository to a vast audience throughout in the Arab region and throughout the world

B. The National Theses Library

For this purpose the National Thesis Library was created in 1992 to be part of the Ain Shams University Information Network (ASUNET). The intention was to store all theses submitted at Egyptian universities to obtain Masters and Doctoral scientific grades and translations and/or resumes of thesis of Egyptian scholars qualifying abroad, who got a degree-equivalence from the Supreme Council of Universities. All subjects of natural science and humanities were considered.

Due to the lack of sufficient resources, only half the influx of documents was converted to the corresponding e-form. Also: Faculties and Universities throughout the country differ in acceptance of the idea so that there is still some important leak of information and inter-Universities communications problems. The description of the dual conversion system by ASUNET is shown in Fig. 1.

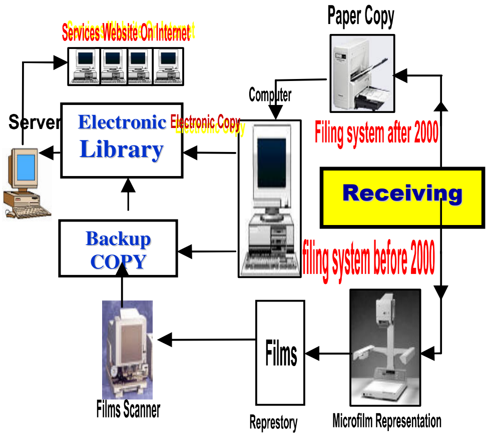

Fig. 1. The conversion system (ASUNET diagram)

The uncertainty concerning Egyptian copyright laws, which are still not sufficiently updated to consider e-publishing issues, made it obligatory to sign with each and every researcher special agreements concerning providing his or her thesis to online users. Table I shows an example of the rate of delivering theses from Egyptian universities.

It is therefore only logical that the intended project is conceived to contain the following subsequent steps:

1. Converting existing contents

   In this first stage all exiting Microfilms have to be converted by means of special scanning devices and software programs to e-files. The duration of this stage is estimated by 12 months.

2. Enabling stage

   After the completion of the first stage, the RDBMS responsible for structure and retrieval of the

TABLE I
RATE OF THESES DELIVERED TO ASUNET FROM 1/1/2006 TO 24/8/2006

UNIVERSITY	**NO. OF THESES**
CAIRO	2127
AIN SHAMS	1360
EL MENOFIA	810
SUEZ CANAL	500
ALEX	410
ASSUIT	400
DAMISCS	300
EL MINIA	118
TANTA	100
EL MANSORA	80
SOUTH OF VALLY	75
TOTAL	**6280**

generated documents is created. All access points, either through the internet or through other offline sources, have to be correctly defined and realized. This is estimated to cost 6 months of intensive development effort.

C. Applying Security Measures

Security schemes and measures are then applied so that only permitted (or paying) users gain access to the information body while other visitors can only see the abstracts of scientific documents. In this stage an appropriate digital-rights- management scheme has to be applied. 2-3 months are expected for this stage

D. Cooperation with internal and external sources

In the last stage of the project regional and international partners are sought so that open accessibility to interested parties world-wide is guaranteed. Additional 3 months are estimated.

E. Intellectual Property and legal, Managerial Issues

According to the current Egyptian law, copyrights of electronic theses and dissertations remain with the author; however, students assign rights to publish the electronic version online. The accessibility of digital documents distinguishes them from their print counterparts, and complicates their decision. The digital equivalent of practices that were common with print material may place the Owners at unacceptable risks, merely because they are more visible and may have greater consequences. Selection of a central deposit must begin with an understanding of the current uncertainty in the application of copyright to digital resources. Copyright law, still written on and, most cogently about paper, is in flux as it extends to digitized materials.

Evidence of the copyright status and documentation of efforts to obtain permission to make copyrighted digital resources available—including a signed copyright waiver from the copyright holder or written documentation that details a good faith effort to secure such permission—are required as part of the deposit process. Under special circumstances, digitized material that is copyright-protected, but which will fall into public domain within a short time frame, may also be considered for deposit [16].

Note that copyright may cover software use as well as digital content. In addition to copyright, privacy and donor restrictions must be considered. It is the depositing unit's responsibility to ensure that these rights are not breached by the digitization and use of such materials. On comparison with other ETD's, Table II shows a survey [6-15], [24] on some feature of asset of ETD's databases.

V. MODERN STANDARDS OF SCIENTIFIC INFORMATION RETRIEVAL

Since the early years of the 21st century much emphasis has been put on redefining the ordinary structured machine search to incorporate more and more intelligence and thus be able to tackle with unstructured data. This is not only due to the fact, that the world wide web became the number one way of information retrieval, but also due to the unprecedented explosion of contents which made discovering information a much more challenging technological task than ever imagined, a task needing exploration of all aspects of information management and access and application of expertise in a wide variety of topics, including digital libraries, natural-language understanding, statistics, computer science, hypertext, etc.

Research in the past years has concentrated on the following information retrieval (IR) issues:

A. Domain-specific meta-data information (so-called ontologies) for key-word mapping and classification (summarized from [17], [18], [19] & [20]):

Modern search systems largely use keyword-based algorithms, which is known to be the most effective and efficient method for practical, general-purpose search. In the Web context, this keyword indexing has been enhanced by deriving indexing information from link structures. However, external meta-data (collected in catalogs) supplied by authors or third parties may also be used. Tools for searching these catalogs have developed through the years and depend largely on well-structured catalog records that include controlled subject vocabularies and name authorities.

TABLE II
A SURVEY ON SOME ETDs DATABASES

Library or consortium	**No. or Thesis**	**Digital thesis collection only/since**	**Full-text**	**Place**	**Comments**
Melvyl catalog	250,000,000	No/ No Info	Yes	N. America	Full Text available for subscribers, students and university staff **http://www.lib.berkeley.edu/** http://www.cdlib.org/
The Center for Research Libraries	433000	Yes/ No Info	No	N. America	Full texts are available upon request in the Sub-Systems of the respective contributing libraries . **http://www.crl.edu/**
Hollis Catalog	-	No/ No Info	-	N. America	There are over 90 libraries at Harvard University with extensive collections in numerous subject areas. Each library has its own set of policies and procedures and should be consulted for specific questions and for assistance with research. http://lib.harvard.edu/catalogs/hollis.html
West Virginia University	3000	Yes/1997	Yes	N. America	Restricted access to subscribed users. http://www.wvu.edu/~thesis/
Tesionline.com	10557	Yes/2000	Yes	Europe	Commercial Provider. Full texts for paying subscribers only. http://www.tesionline.com/intl/index.jsp
DATAD (Database of African Thesis and Dissertations)	-	No/2000	No	Africa	http://www.aau.org/datad
National Lib. of Wales	15000	No/1984	No	Europe	Web Database still being developed

The most notable product of this work is the Z39.50 protocol, a US National Information Standards Organization (NISO) standard that nearly all Library Management System vendors support. As search has evolved, however, to be the most intensive activity on the Web, commercial interests have made it hard to use meta-data directly in search algorithms. This is mainly due to so called “Web spamming”, i.e. providing misleading meta-information to search engines. Therefore, trusted sources of meta-data, such as a libraries, publishers, or institutional repositories, are considered to be very helpful in this respect. The Open Archives Initiative Protocol for Meta-data Harvesting (OAI-PMH) is the most widely used protocol for this purpose. It is designed to let service providers access any type of meta-data written in XML format and related to any form of information object. OAI-PMH provides thus access for search engines and their crawlers to information from the “depth” of the Web, such as scientific databases or publishers' repositories [18].

An ontology is generally regarded as an abstraction and formal description of the above, loose concept of meta-data. It is a “designed artifact consisting of a specific shared vocabulary used to describe entities in some domain of interest, as well as a set of assumptions about the intended meaning of the terms in the vocabulary”. Another definition of ontology is “an explicit specification of a conceptualization”. This definition is used in Artificial Intelligence and is concerned with the formal symbolic representation of knowledge. Fundamental to this approach is the notion of conceptualization. That is, an abstract and simplified view of the world, or domain of interest, which is being represented. A conceptualization consists of objects or entities that are assumed to exist in the domain of interest as well as the relationships that hold between them. Relationships can be interpreted as roles. The set of objects about which knowledge is being expressed is referred to as the universe of discourse. The universe of discourse and the relationships that hold in it are expressed in a declarative formal vocabulary constituting the knowledge about a domain. It is stated that the task of intelligent systems, in general, is to formally represent and commit to some conceptualization implicitly or explicitly. An explicit specification of such a conceptualization is called an ontology ([17], [19]).

The ontology of a shared domain can be described by defining a set of representational terms. These terms (lexical references) are associated with entities (non lexical referents) in the universe of discourse. Formal axioms may be introduced to constrain their interpretation and well-formed use. In this respect, an ontology can be seen as the explicit statement of a logical theory. Although such ontologies often assume the form of a taxonomic class hierarchy, they are not restricted to hierarchies. In order for knowledge to be shared amongst agents, agreement must exist on the topics about which are being communicated. This raises the issue of ontological commitment described as “the agreements about the objects and relations being talked about among agents”. A common ontology defines thus the vocabulary with which queries and assertions are exchanged among agents, thereby providing the means to bridge the semantic gap that exists between the lexical representations of information and its non-lexical conceptualization. It is important to note that even though commitment to a common ontology is a guarantee of consistency, it does not always guarantee completeness with respect to queries and assertions made, unless a constructive, deductive definition is introduced allowing the correct classification of much more lexical objects than originally physically listed [20].

b. Intelligent Search engines

Modern commercial search engines use in general a five step procedure to retrieve and display results. This procedure comprises the following [21].

1) Seed list creation
2) Focused crawling
3) Classification & Maintenance
4) Query
5) Ranking

A seed list is a set of URLs and links relevant to the process of IR. Usually seed lists are created using link analysis (automatically), agreements with publishers and employees and webmasters feedback. "Crawlers" are robots reading texts found on sites and processing them according to technical specifications. They are guided by entries in the seed list and don't – in general- assume any other sources of information. The general procedures implemented by those robots is fairly simple: Documents are prioritized tracking the rules set by webmasters, collected documents are sent to the general indexing instance and copies stored so that the software can show parts of the documents containing the query term(s). This is done on independent machine nodes, sometimes called "crawler farms". Each node is assigned a segment of the Web to crawl.

In addition to crawling, loading from specific open sources is usually the most valuable asset in any commercial engine. The OAI discussed above has made finding such sources and getting the desired information easier than before. Also: Most known vendors go into agreements with publishers like ScienceDirect and BioMed which provide access to their up-to-date repositories.

After crawlers gather all the relevant pages and put them into a "working" index every word is read and examined. The classification process is then started usually following two different schemes: Subject classification and type classification. Subject classification uses dictionaries and ontologies as mentioned above. The more meta-data descriptions are intelligible, the more results tend to converge towards the originally intended query. Type classification shows users profiles of pages (i.e. flags containing remarks as "homepages", "publishing lists" etc). Once the classification is complete, indexes are searched and results ranked in an appropriate order. Vendors differ in their Web-page sources from university-pages to private collections of articles, and from free downloadable contents to paying sites.

Query transformations or "intelligent query rewriters" are often used to optimize user requests and remove non-essential search phrases. Many search engines enable users to logically narrow or broaden their requests. Ranking depends mostly on term values, location and frequency of occurrences within documents. Global frequencies of terms within the whole index are also taken into consideration. Link analysis is a relevant part in any valid ranking system. Link values may be calculated by counting the number of links to a given page. The more links a page has, the higher its rank. Statistic scores are used to compensate for pages loaded directly from accessed databases. Finally: Large scientific dictionaries may be created with time if appropriate link analysis techniques are applied comprising more and more scientific terms used uniformly throughout the whole search space.

C. Heterogeneous data sources, Parallel, distributed architectures, Partitioning and Scientific Data Grids (from [21] & [22] & [23])

Scientific database applications may use multi-terabyte datasets especially in fields such as astronomy and biology. They are therefore particularly suited for the application of parallel, distributed architectures, dataset-portioning and/or automated physical design techniques. Usual automated physical design tools focus on the selection of indexes and materialized views. In large-scale scientific databases, however, the data volume and the continuous insertion of new data allows for only limited indexes and materialized views. By contrast, data partitioning and/or distributed work-load allocation does not replicate data, thereby reducing space requirements and minimizing update overhead. It is common to define the overall infrastructure of the IR system as a SDG (Science Data Grid).

One useful definition of a SDG, leaning on IBMs explanation of the term (see [25]) may be: An SDG is scientific data set using a set of open standards and protocols, to gain access to applications and data, processing power, storage capacity and a vast array of other computing resources over the Internet. An SDG is a type of parallel and distributed system that enables the sharing, selection, and aggregation of resources distributed across multiple scientific sites based on the resources availability, capacity, performance, cost and users' quality-of-service requirements determined by the contributing scientific bodies.

VI. ISSUES CONCERNING ARABIC ONTOLOGIES

Recent world-wide research has focused on the issue of building lexical resources called WordNets whose functionalities enable lexical level translation from and to dozens of languages. Arabic is still missing and the CIA[1] funded project [27] describes a method of construction of an Arabic WordNet conform to the universally accepted PWN 2.0 (Princton WordNet) standard. The method also suggests relying upon an Arabic-language extension to the SUMO (Suggested Upper Merged Ontology)[2], in which word meanings are defined with a machine understandable semantics in first order logics. The ASUMO concepts are seen as Interlingual Indexes (ILIs) connecting different wordnets together. In [26] an automatic, lexeme-level bootstrapping method is proposed which depends on Arabic-English Corpora and an English WordNet. It is argued that a lexeme-based approach to the construction of Arabic ontologies is more appropriate than a root-based one. In any case efforts in the direction of constructing Arabic ontologies are deemed to be essential at this point of time [26], [27].

VII. CONCLUSIONS

Our conclusions can be viewed as a set of recommendations to apply in the course of physical implementation of the ETD project at ASUNET. We hope to be able to fulfill our initial goal of providing the Egyptian universities with a modern and powerful tool to access scientific information if the following requirements are met:

1. Creating a professional business plan for the creation of the said ETD including financial, technical and managerial milestones to be met in a realistic time-frame.
2. Initiating an inter-disciplinary group concerned with the issues of E-Copyrights with the aim to solve them to be able to provide accessibility to FULL-Text documents in a legal, organized way.
3. Technical Recommendations :
    i. Adopting the above described ETDMS standard for metadata-specification and the OAI-PMH Protocol.
    ii. Constructing intelligent search engines using the above described five-step search procedures and ontologies on top of their usual search algorithms. This implies creating scientific interest within the Egyptian universities in the general theme of electronic search by means of ontologies and in the issue of “Arabic enabled ontologies” specifically and commencing R&D in this direction. We strongly recommend adopting the WorldNet initiatives
    iii. Redefining the overall infrastructure of the IR system as a SDG (Science Data Grid), open-ended and able to incorporate future scientific data-management requirements

REFERENCES

[1] Sayeed Choudhury , Eva Müller "Enriching E-theses", *A briefing paper for the workshop on e-theses*, Amsterdam, January 2006

[2] John MacColl, "Electronic Theses and Dissertations: a Strategy for the UK", *Ariadne Issue 32*

[3] Fox, E. "Contribution by Edward A. Fox regarding Networked Digital Library of Theses and Dissertations (NDLTD) for UNESCO Meeting", September 27-28, Paris, 1999.

[4] Atkins, Anthony, Edward A. Fox, Robert France and Hussein Suleman (editors). 2001. "ETD-ms: an

[1] US Central Intelligence Agency

[2] Is a freely available, formal ontology of about 1000 terms and 4000 definitional statements. It is provided in a first order logic language called Standard Upper Ontology Knowledge Interchange format and also translated into the OWL semantic web language. It has been subjected to formal verification with an automated theorem prover and extended with a number of domain ontologies, which are also public, that together number some 20,000 terms and 60,000 axioms. SUMO has been mapped by hand to the WN lexicon of 100,000 noun, verb, adjective and adverb senses, which not only acts as a check on coverage and completeness, but also provides a basis for application to natural language understanding tasks. SUMO covers areas of knowledge such as temporal and spatial representation, units and measures, processes, events, actions, and obligations. Domain specific ontologies extend and reuse SUMO in the areas of finance and investment, country almanac information, terrain modeling, distributed computing, endangered languages description, biological viruses, engineering devices, weather and a number of military applications.

Interoperability Metadata Standard for Electronic Theses and Dissertations -- version 1.00". Available from http://www.ndltd.org/standards/metadata/ETD-ms-v1.00.html. 2001.

[5] DCMI. 1999. "Dublin Core Metadata Element Se"t, Version 1.1: Reference Description. Available from http://www.dublincore.org/documents/dces/, 1999.

[6] Fox, Edward A. 2000. "Core Research for the Networked University Digital Library (NUDL)", NSF IIS-9986089 (SGER), Project director, E. Fox., 5/15/2000 - 3/1/2002.

[7] Fox, Edward A., John L. Eaton, Gail McMillan, Neill A. Kipp, Laura Weiss, Emilio Arce, and Scott Guyer. 1996. "National Digital Library of Theses and Dissertations: A Scalable and Sustainable Approach to Unlock University Resources", *D-Lib Magazine*, September 1996. Available at http://www.dlib.org/dlib/september96/theses/09fox.html.

[8] Fox, Edward A., Brian DeVane, John L. Eaton, Neill A. Kipp, Paul Mather, Tim McGonigle, Gail McMillan, and William Schweiker. 1997. "Networked Digital Library of Theses and Dissertations: An International Effort Unlocking University Resources", D-Lib Magazine, September 1997. Available at http://www.dlib.org/dlib/september97/theses/09fox.html.

[9] Fox, Edward A., Royca Zia, and Eberhard Hilf. 2000. "Open Archives: Distributed services for physicists and graduate students (OAD)", NSF IIS-0086227, 9/1/2000-8/31/2003. Project director, E. Fox (w. Royce Zia, Physics, VT, and E. Hilf, U. Oldenburg, PI on matching German DFG project).

[10] Fox, Edward A., J. Alfredo Sánchez, and David Garza-Salazar. 2001. "High Performance Interoperable Digital Libraries in the Open Archives Initiative", NSF IIS-0080748, 3/1/2001-2/28/2003. Project director, E. Fox (with co-PIs J.Alfredo Sánchez, Universidad de las Américas-Puebla --- UDLA, and David Garza-Salazar, Monterrey Technology Institute --- ITESM, both funded by CONACyT in Mexico).

[11] Kipp, Neill, Edward A. Fox, Gail McMillan, and John L. Eaton. 1999. "FIPSE Final Repor"t, 11/30/99. Available from http://www.ndltd.org/pubs/FIPSEfr.pdf (PDF version) and http://www.ndltd.org/pubs/FIPSEfr.doc (MS-Word version).

[12] Lagoze, Carl and Herbert Van de Sompel. 2001. "The Open Archives Initiative Protocol for Metadata Harvesting". *Open Archives Initiative*. January 2001

[13] Library of Congress. "Program for Cooperative Cataloguing Name Authority Component Home Page", 2001.

[14] NDLTD. 1999. "Publishers and the NDLTD". *NDLTD*, July 1999. Available from http://www.ndltd.org/publshrs/.

[15] Powell, James and Edward A. Fox. 1998. "Multilingual Federated Searching Across Heterogeneous Collections", *D-Lib Magazine*, September 1998. Available at http://www.dlib.org/dlib/september98/powell/09powell.html.

[16] OCLC. 2001. Persistent URL Home Page. Dublin, "OH: OCLC Online Computer Library Center", 2001. Available from http://purl.oclc.org/.

[17] Data Description and Archives for Scientific Research in the future, Imaging and Media Lab, University of Basel, Switzerland, *Conference Proposal for IS&T Archiving* 2006.

[18] Carl Lagoze and Amit Singhal, "Information Discovery: Needles and Haystacks", *IEEE Internet Computing*, vol. 9, no. 3, pp. 16-18, 2005.

[19] Pretorius, A.J., “Lexon Visualisation : Visualising Binary Fact Types in Ontology Bases”, *Chapter 2, MSc Thesis, Brussels, Vrije Universiteit Brussel*, 2004.

[20] Elnaser Abdelwahab, „Entwicklung eines zentralen Data-Dictionary fuer ein Tumordokumentationssystem“, *PHD thesis, Giessen, University of Giessen*, 1997

[21] "How Scirus works", *Technical White paper*, August 2004, http://www.scirus.com/press/pdf/WhitePaper_Scirus.pdf#search=%22How%20Scirus%20works%22

[22] Stratos Papadomanolakis Anastassia Ailamaki, "AutoPart: Automating Schema Design for Large Scientific Databases Using Data Partitioning", *Internal Report: CMU-CS-03-159, Computer Science Department School of Computer Science, Carnegie Mellon University.*

[23] Kai Nan, Deting Yang, "Requirements and Architecuture for Data Grid Middleware", *Computer Network Information Center, Chineese Academy of Sciences*, PNC 2003 Bangok, PP-Presentation, 2003.

[24] Ian Foster, "What is the Grid? A three point check-list", *Argonne National Laboratory & University of Chicago*, 2002

[25] IBM Solutions Grid for Business Partners, "Helping IBM Business Partners to Grid-enable applications for the next phase of e-business on demand", *Grid-Computing series*, 2002

[26] Mona Diab. "The Feasibility of Bootstrapping an Arabic WordNet Leveraging Parallel Corpora and an English WordNet". *In Proceedings of the Conference on Arabic Language Resources and Tools, Cairo, Egypt*, September 2004.

[27] Fellbaum C., M. Alkhalifa, W. Black, S. Elkateb, A. Pease, H. Rodriguez, P. Vossen 2006 “Introducing the Arabic WordNet project”, *In Proceedings of the 3rd Global Wordnet Conference, Jeju Island, Korea, South Jeju*, Januaru 22-26, 2006.