Retrieval of Arabic Articles
The Case of Nahda Journals

FATMEH CHARAFEDDINE
RESEARCH SERVICES HEAD OF DEPARTMENT
UNIVERSITY LIBRARIES- AMERICAN UNIVERSITY OF BEIRUT

ORIENT INSTITUT OF BEIRUT
SYMPOSIUM
HIDDEN TREASURES – JOURNALS OF THE MIDDLE EAST
THURSDAY, 20 OCTOBER, 10:00-16:00, 2011
LIBRARY OF THE ORIENT-INSTITUT BEIRUT

Current Situation

- A big demand for Arabic periodical literature.
- Rich collections of Arabic periodicals in university and special libraries with little discovery tools.
- Very few indexing services, most of which are short lived lack comprehensiveness, or cover individual titles.
- Very few databases covering Arabic periodicals literature.
- Uncoordinated Arabic digitization and problems with Arabic OCR.

Significance of al-Nahda

- Cultural and political awakening that started early in the 19th century, at the downfall of the Ottoman Empire.
- Connected to the cultural shock of the East-West encounter, best manifested after Napoleon's invasion of Egypt in 1798.
- Famous for its intellectual discourse expressing the need for modernization and reformulation of Islamic doctrine.
- Major political changes (fall of Ottoman Empire, World War I, mandates, independence, creation of Israel).
- Birth of Arab nationalism.

Why Nahda Press ?

- Prevailing medium to publish, exchange and debate emerging ideologies and ideas.
- Reflected the reformist spirit and the political mood of the time in its various views, all over the Arab world.
- Great prominence of its writers in modern Arab thought, literature and culture (Bustani, Yaziji, Shidyaq, M. Abdo, R. Rida ...)
- Introduced modern concepts and ways of thinking promoting progress following the European model.
- Proliferation, good circulation record and long runs of publishing.

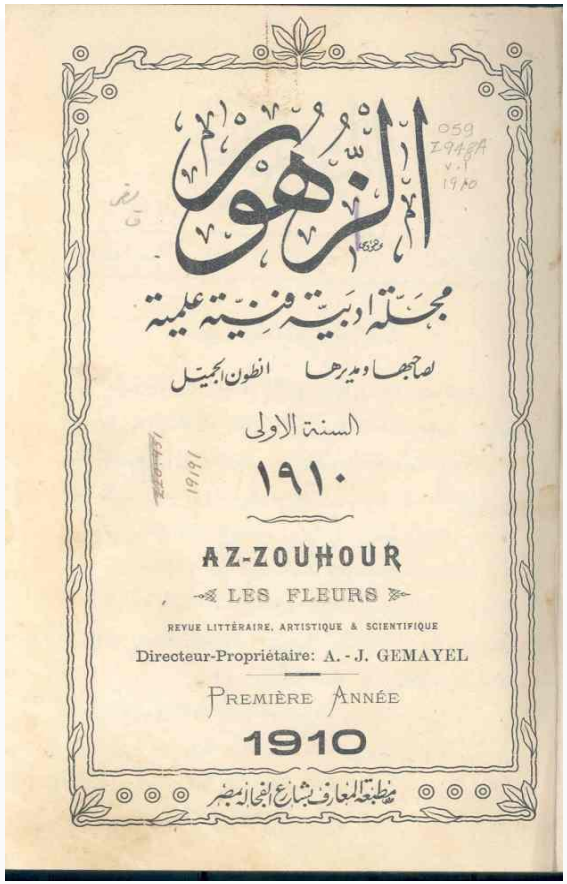
Press Profile

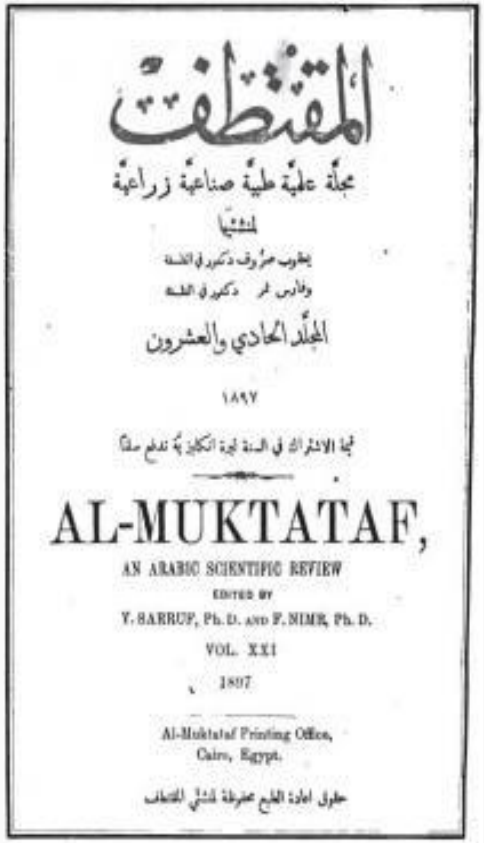
- Most Nahda journals were published independently by intellectuals from the educated elite.
- Publication started early in the 19th century and proliferated during the 1870s and thereafter.
- Number of titles covered by de Tarazi (تاريخ الصحافة العربية) came to 3,023 in total, 1,072 journals and 1,951 daily newspapers with Egypt ranking first, 543 titles.
- Journals were with a reformist goal, often stated in slogans on the cover and title page of the journal.
- No clear specialization, a holistic coverage of timely topics by most journals and daily newspapers

Highlights

- Prevalence of themes mostly related to modernism : education, scientific knowledge, technology, political liberation, revision of Islam vs. liberation and rejection of the past, nationalism, and women's rights.
- Appearance of pioneer works in the realm of modern Arab thought and literary creation (novel, play…).
- Included female magazine editors, which started to appear at the turn of the 20th century.
- Contributed to the forging of new vocabulary capable of stating incoming ideas and concepts.


السنة الاولى
مجلة علمية ادبية تاريخية
REVUE DE HAWDATT EL-MA'ARIF


الزهور
مجلة ادبية فنية علمية
لصاحبها ومديرها انطون الجميل
السنة الاولى
١٩١٠
AZ-ZOUHOUR
LES FLEURS
REVUE LITTÉRAIRE, ARTISTIQUE & SCIENTIFIQUE
Directeur-Propriétaire: A. - J. GEMAYEL
PREMIÈRE ANNÉE
1910
مطبعة المعارف بشارع الفجالة بمصر


المقتطف
مجلة علمية طبية صناعية زراعية
المجلد الحادي والعشرون
١٨٩٧
AL-MUKTATAF,
AN ARABIC SCIENTIFIC REVIEW
EDITED BY
Y. HARRUF, Ph. D. AND F. NIMR, Ph. D.
VOL. XXI
1897
Al-Muktataf Printing Office,
Cairo, Egypt.

Single Nahda Journal Indexes

الأديب البير أديب 1942-1967
فهرست مجلة الاديب (1942-1967) اعداد مريم فران هاشم ؛ تحت اشراف جبور عبد النور

مجلة الثقافة (القاهرة)1939 -1952
دراسة وفهرسة مجلة الثقافة اعداد هلال مصطفى الناتوت

الجنان المعلم بطرس البستاني 1870-1886
فهرست مجلة الجنان إعداد جبران أيوب إشراف جبور عبد النور

الرسالة: مجلة اسبوعية (االقاهرة)أحمد حسن الزيات 1933-1953
فهارس "الرسالة" رسالة مقدمة من ناهية جبرائيل ايوب، اسعد توفيق حرفوش؛ تحت اشراف جبور عبد النور

الضياء (القاهرة) ناصيف اليازجي 1898-1906
مجلة الضياء فهارس تحليلية مع مقدمة اعداد سلوى فؤاد حماد؛ اشراف محمد يوسف نجم

العرفان (صيدا) أحمد عارف الزين 1909-
فهارس العرفان، رسالة لنيل شهادة الكفاءة في الادب العربي : جلول, فاطمة

المشرق لويس شيخو 1898-
فهارس المشرق العامة

المعرض ميشال زكور1921 - 1936
فهرس مجلة المعرض سلمى عزيز سليمان ؛ باشراف جبور عبد النور

المعلم الجديد (بغداد) 1935-1979
كشاف "موضوعي" بمجلة المعلم الجديد اعداد نورهان عبدي رشيد؛ مراجعة جاسم جرجيس.

المقتبس محمد كرد علي 1906-1917
فهارس المقتبس اعدها رياض عبد الحميد مراد : مراد, رياض عبد الحميد

المقتطف (القاهرة) يعقوب صروف و فارس نمر 1876- 1952
فهرس المقتطف، 1876-1952 الدراسات العربية في الجامعة الاميركية

المكشوف : جريدة اسبوعية فؤاد حبيش 1934-1944
فهرست مجلة المكشوف عايدة خوري، باشراف جبور عبد النور

المنار محمد رشيد رضا 1898-1935
فهرس مجلة المنار، 1898- 1935 اعداد كوسوجي ياسوشي، يوسف حسين ايبش، يوسف قزما خوري

الهلال (القاهرة) جرجي زيدان 1892 -1985
كشاف الهلال مجلة الآداب و الفنون و العلوم رضا صبحى الوزيري

صحيفة ثمرات الفنون عبد القادر القباني 1875-1908
صحيفة ثمرات الفنون، "فهرسة و دراسة" هدى صياح ؛ بإشراف جبور عبدالنور،

مينرفا مجلة ادب وفن واجتماع ماري يني 1923 - 1927
مجلة "مينرفا" دراسة وفهرسة اعداد ماري هدايا ؛ باشراف اسامة عانوتي

Collective Arabic Journal Indexes

- محتويات الدوريات العربية : مجلة فصلية توثيقية Serial/قبرص: مؤسسة دراسات الشرق الاوسط (سام منسى)
  TOC subject and author index 1994-1998
- محتويات الدوريات Serial/ تونس : جامعة الدول العربية Only TOC 1988-89
- كشاف الدوريات العربية، 1876-1984 / اعداد عبد الجبار عبد الرحمن 1989
  top-down chapter for each subject
- الكشاف التحليلي للصحف و المجلات العربية Serial / القاهرة: لجنة الفهارس العربية , 1962-1967(شعبان خليفة)
  تركيز على الجمهورية العربية (مصر).
- الفهرست Serial/ رئيس التحرير ميشال نوفل بيروت 1981-1989 (قائمة رؤوس الموضوعات العربية, سويدان) (سام منسى)

- Ceased and short lived serial publications.
- Didn't follow standard subjects list or indexing structure and form.
- Few specialized indexes, subject (7 in AUB UL), country (Syria, Jordan Iraq).
- No representation of Nahda journals.

Middle East Studies Databases

- **Al-Nahar Online:** keywords between 1988 -1994, full-text search 1995-
- **AskZad** includes **AskZad Full-Page Newspapers** *top 100 Arabic language news publications from the Middle East for presentation as complete archives*. **Pan-Arab Academic Journal Index** *indices of 700 journals published in Egypt and the Arab world*. **Pan-Arab Academic Journals** *is a full-content database of 191 Applied Science and Social Science Academic Journals published in the Middle East*. ***Pan-Arab Peer-Review Articles*** *cover research published in credible, but non-Academic, journals and periodicals. Subject classification and searching are in Arabic.*
- **EduSearch** index and full text of 240 journals in the field of education 1940-. Subject classification and searching are in Arabic.
- **Index Islamicus**
  Produced the School of Oriental and African Studies in London University. Covers over 3,000 journals together with conference proceedings, and monographs in the field of Middle East and Islamic studies. No Arabic coverage.
  Middle Eastern & Central Asian Studies (MECAS) includes the following: *Middle East Bibliography* (1946 - 2001), *Middle East Book Bibliographies, Theses & Dissertations, MECAS Citations Database* and *School of Oriental & African Studies* (SOAS) *Library Catalogue* (1900 - present).
- **Multidataonline:** covers 300 Arabic sources, and includes **Index Arabicus,** the only database covering titles from Nahda. Subject classification and searching are in Arabic.

INDEX ARABICUS

Online bibliography of Arabic periodical articles covering the period, 1870-1969. Modeled on Index Islamicus, it was originally compiled on cards by members of MELCOM-UK in the 1970s, sent to Beirut for printing, and apparently lost in the turmoil of the Lebanese civil war.

More recently the cards came into the possession of the University of Imam al-Ouzai in Beirut, who have entered their contents into an online database, provided by Multidata Services.

ادبي : مجلة شهرية ادبية	ضياء
الأديب	العرفان
البيان	فتاة الشرق
الجامعة	الكاتب المصري
الجامعة العثمانية	المشرق
الجنان	المقتبس
الرسالة	المقتطف
الرواية	المنار
الزهور	الهلال

Searching al-Nahda Journals

- Lack of systematic and comprehensive indexing services.
- Most published print indexes are one journal index which in most cases lists articles under broad subject categories, and provide author indexes.
- Most collective journal indexes don't cover the Nahda period.
- The few databases of Arabic periodicals are mostly concerned with current periodicals literature (AskZad, Multidataonline, EduSearch).
- The only article database specialized in Middle Eastern Studies is Index Islamicus which doesn't cover Arabic publications.

Arabic Indexing Tools

- Most Arabic indexing tools are subject heading lists and few thesauri that are top level; they describe general concepts that are common across domains.
- Revising and updating is not on-going. Some are never updated.
- Lack comprehensiveness (Khazindar, 1994) covered 10,000 subject heading vs. LCSH quarter of a million (1997); Khazindar started in 1958 with 2900 headings, LC started 1898.
- There are some Arabic subject thesauri, many are translations from thesauri published by international agencies such as UNESCO, UNBIS.

General Subject Heading Lists and Thesauri

رؤوس الموضوعات العربية إشراف ناصر محمد السويدان الرياض : جامعة الملك سعود1978, 1985

قائمة رؤوس الموضوعات العربية = Arabic subject headings list الجامعة الاميركية في بيروت، مكتبة يافث التذكارية، دائرة بيروت, 2000

قائمة رؤوس الموضوعات العربية الكبرى شعبان عبدالعزيز خليفة، محمد عوض العايدي.القاهرة : المكتبة الأكاديمية 1981, 1994-2000ملحق

قائمة رؤوس الموضوعات العربية الموحدة : إتيم , محمود أحمد تونس: المنظمة العربية للتربية والثقافة والعلوم, 1995

قائمة رووس الموضوعات العربية إعداد إبراهيم أحمد الخازندار. الكويت: جامعة الكويت 1958 , 1983

الجامعة مكنز ثلاثي اللغات ـ العربية، الإنجليزية، الفرنسية UNBIS تونس: جامعة الدول العربية, 1993

= المكنز الموسع = Expanded thesaurus Thesurus etendu عمان: مؤسسة عبد الحميد شومان , 2001 CD_ROM

Expanded Thesaurus المكنز الموسع

Expanded thesaurus = Thesurus etenduالمكنز الموسع =

عمان: مؤسسة عبد الحميد شومان , CD_ROM 2001

- The largest Arabic tri-lingual thesaurus. It contains 50000 terms and 25000 descriptors within 27 facets in Arabic, English, and French.
- The ability to search thru Key-Words-Out Of-Context (KWOC) List using any of the three languages.
- It covers modern development issues (science 4770 agriculture 2430, health 1366, Islam 748 …)
- A new and revised edition is under construction.

DomainThesauri

- قائمة رؤوس الموضوعات العربية في العلوم الاجتماعية 1975
- المكنز الشامل للمصطلحات في مجال التنمية 1979
- مكنز مصطلحات علم المكتبات والمعلومات 1980
- المكنز الاسلامي, المدينة, الجامعة الاسلامية 1983
- مكنز العمل 1989 .
- المكنز العربي لعلوم الأرض 1988 .
- المكنز السكاني متعدد اللغات 1988 .
- مكنز المياه الدولي 1990 .
- مكنز اجروفوك 1993 .
- مكنز التربية والثقافة والعلوم 1994-1995 .
- مكنز الفيصل :شامل في علوم الحضارة 1994 .
- مكنز الأرشيف ط.2 CD-R1997 .
- مكنز الفولكلور 2005
- مكنز البنك الإسلامي للتنمية 1992
- مكنز الوقف 2005
- المكنز التربوي 2008

- Social Sciences
- Development (OECD)
- Library sciences
- Islam
- Labor
- Earth sciences
- Population
- Water resources
- Agriculture (AGROVOC)
- Education, culture and sciences (UNESCO)
- Civilization
- Archives
- Folk arts
- Islamic Bank Development
- Endowment
- Education (UNESCO-IBE)

Needed Solution

- Index Nahda journals using controlled and structured English-Arabic vocabulary system (thesaurus or ontology) covering concepts, ideas, people, organizations, events, places, and significant dates of the period together with their different types and levels of relationships.

- Introduce semantically enhanced search to improve precision and allow for linking between concepts, clustering, and facet searching.

- Use delivery application which offers advanced search, browse and navigate to different levels of the hierarchy of a journal.

Why Controlled and Structured Vocabulary ?

- Thesaurus based indexing versus free text based indexing (authors words found within the document).
- Controlled vocabularies provide better search results in terms of speed in precision and referencing among terms. Structured vocabularies allow for better terminology selection in search limiting and expansion.
- The use of thesauri electronically could be very effective. Linking, alternative term searching, results clustering, and search faceting can become automatic.

Thesaurus Principles

- Major principles in the design and development of a thesaurus :
  - Eliminating lexical ambiguity
  - Controlling Synonyms
  - Establish relationships among used terms
  - Facet analysis
- For our purpose a thesaurus has to deal with retrieval problem specific to the Arabic language.
- Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies ANSI/NISO Z39.19-2005

Lexical Ambiguity

- Ambiguity occurs in natural language with homographs, when a word has more than one meaning باب، صوت، طقس، رمضان

- A controlled vocabulary must compensate for the problems caused by ambiguity by using parenthetical qualifiers to ensure that each term has one and only meaning:
  - Mercury (planet),
  - Mercury (metal),
  - Mercury (mythology).

Selected Semantic Relationships

Equivalency (USE, UF)

Synonymy UN / United Nations
Lexical variants pediatrics / paediatrics
Near synonymy smoothness / roughness

Hierarchy (BTG, NTI,BTP, NTP)

Generic or IsA birds / parrots
Instance or IsA sea / Mediterranean Sea
Whole /Part brain / brain stem

Associative (RT)

- Cause /Effect accident / injury
- Process /Agent fire/ flame
- Process /Counter-agent fire/flame retardant
- Action / Product writing / publication
- Action /Property driving / speed
- Action /Target teaching / student
- Object/Origins water / well
- Raw material /Product grapes / wine
- Field / Object neonatology / infant

Facet Analysis

- Facets may be defined as common attributes corresponding to the article and are helpful in retrieving it.
- Attributes that might be selected as facets for content objects are:
  - Author affiliation -(place of birth, political, intellectual …)
  - Big historic events (wars, visits, conferences, missions …)
  - Document type – (article creative writing, translation, image…)
  - Geographic details (rivers, mountains, street names …)
  - Date periods (wars, mandates, …)
  - Organizations (political parties, organizations, schools …)
  - Top topic – the top subject of the content object

Arabic Language Ambiguities

- In thesaurus building there are some problems specific to the Arabic language.
- Ambiguity of term without their vowel diacritics. The word بر could be read as bar (land), barra (was dutiful), birr (kindness).
- Arabic is highly derivational, diacritics are very important to indentify the various derivational patterns درس could be read he studied, he taught and lesson, depending on the diacritic on the middle letter سكون، شدة، فتحة
- No capital letters to identify, proper nouns could be confused with regular ones جمال.
- Interchangeable use of some letter أ ا ي ى ه ة

Middle East Studies Thesaurus

"since there is no standard subject heading list or thesaurus in education, we developed our own list." EduSearch answer to my question Oct. 18, 2011.

- No thesaurus specialized in Middle East Studies.
- It is important to build such a thesaurus using domain specific topic maps, existing thesauri, subject heading lists and subject indexes.
- Retrieval will improve with faceted searching using article attributes such as: author affiliation or place of birth; material type, illustrations, geographic names, time periods, key events, institutions, and key topics (east-west) (Islam-modernization).
- Bilingual controlled vocabulary is required for best research results and end user convenience.

Current efforts

- A lot of research on the automatic processing of Arabic language for the purpose of auto-indexing of Arabic text (root stemming, pattern rhyming, co-occurrence).
- Research is done on solving Arabic OCR problems (cursive writing, change of shape according to position within the word, diacritics, letters with dots, change of shape with different fonts.)
- Several Arabic ontologies are on their way Arab WordNet (based on Princeton WordNet) and Arabic ontology (University of Beirzit). Both are linguistic and not domain specific.
- Ontologies and free text indexing are additional search facilitators but they don't satisfy the need for a unified and consistent way to retrieve concepts and terminology specific to a domain or field of knowledge.

Where to Start

- Don't re-invent the wheel. Select an existing subject list and use it as a list of core base concepts, extend it downward to more specific concepts.
- Define the methodology of developing the thesaurus in terms of structure (listing, hierarchal ..), term form (nouns, verb nouns, phrase, compound terms, plurals), and semantic relationships (UF, BT, NT, RT ...) following set standards and best practices.
- Identify or start name authority lists of people, place, organization and events of the period.
- Consult with subject specialists in choosing preferred terms, writing scope notes and establishing concepts trees.
- Try to partner with other organizations working on same issues.

Opportunity and Challenges

- We are gathered here and we know what we need, and can know more about what is being done.
- The big challenge is who will take the initiative and give the steady commitment to build and maintain such a thesaurus ?

Thank You And Happy Anniversary OIB