

# **Bibliotheca Alexandrina**

## **Online Revolutions in the Arabic World: Event Harvesting in Perspective**

**2011 IIPC General Assembly**

**Dr Magdy Nagi**

*Head of ICT Sector, Bibliotheca Alexandrina*



NATIONAL FLAG ON THE STEPS OF THE LIBRARY OF ALEXANDRIA

# Saving Post-Revolution History (to Disk) for the Future's Access

- Harvesting online content related to January 25, 2011 events in Egypt and possibly similar events elsewhere
- Harvested content goes up in value as time passes
- If it is not done now, it cannot be done later
- The BA has been harvesting:
  - News and politics Web sites
  - Social networking and video/photo sharing sites
  - Satellite channels

# News and Politics Special Web Crawl

- Started February 9, 2011
- 80+ seed URIs
- Seeds collected by asking people what they visit
- 2 daily snapshots
- 4 TB so far
- Yet to be indexed and published

# Crawl Setup

- Heritrix 3 on a single host
- A cronjob is set to start a crawl job at 10:00 AM EET, terminate it after 12 hours, then start over at 10:00 PM EET
- Another cronjob copies the crawled data to two storage hosts and generates the metadata files (.dat.gz)
- Data pools on the crawling host are flushed periodically

# Social Networking and Video/Photo Sharing Web Sites

- 5 sites are searched on certain keywords related to the events
- YouTube video URIs are extracted from HTML and submitted into a queued download service running in the BA's Internet Archive cluster
- The Facebook API is used to enumerate and download videos and images on Facebook pages
- Tweets are collected using both the Twitter API and a specially written Twitter crawler
- Images are enumerated and downloaded using site APIs on both Flickr and Picasa



facebook

twitter 



# Social Networking and Video/Photo Sharing Web Sites



40,000+ videos

facebook.

12,000+ images, 1,000+ videos



600,000+ tweets



4,000+ images



6,000+ images

# Satellite Channels

- 5 satellite channels
- Recorded during scheduled hours daily with focus on capturing news programs
- Recordings are stored on Petabox racks



**TV5MONDE**

Thank You

