"Chorus Vision: Outcome of the Think-Tank"
Why does Search matter

- Because it is everywhere and is useful to the public and to professionals
  - Preferred access means to information on the Internet, Intranet and professional environments.
- And because it generates tremendous economic value

Internet advertising

<table>
<thead>
<tr>
<th>% of 2008 Second-Quarter Revenues</th>
<th>Total – $6.7 Billion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search</td>
<td>44%</td>
</tr>
<tr>
<td>Classifieds</td>
<td>14%</td>
</tr>
<tr>
<td>Sponsorship*</td>
<td>2%</td>
</tr>
<tr>
<td>Banner Ads*</td>
<td>21%</td>
</tr>
<tr>
<td>Display Ad*</td>
<td>9%</td>
</tr>
<tr>
<td>Email</td>
<td>2%</td>
</tr>
<tr>
<td>Rich Media*</td>
<td>7%</td>
</tr>
<tr>
<td>Lead Generation</td>
<td>7%</td>
</tr>
<tr>
<td>Total</td>
<td>100%</td>
</tr>
</tbody>
</table>

Enterprise solutions

<table>
<thead>
<tr>
<th>Forecast (M$)</th>
<th>2006</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>717</td>
<td>1,219</td>
</tr>
</tbody>
</table>

Europe
And audio-visual search

- Internet: the Youtube example
  - More than 59M unique US based visitors monthly
    (source: Compete.com Dec 2007)
  - #2 for search queries in August 08
    - >100% growth year to year.
- A multitude of solutions elsewhere
  - Personalized TV
  - Enterprise video DAM
  - Medical applications
  - Security
  - Mobile/Desktop/etc
- And initial success in the area of delinearized TV.
  - Catch-up TV

<table>
<thead>
<tr>
<th>Search queries (MM) Aug-08</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Expanded Search</td>
<td>17,271</td>
</tr>
<tr>
<td>Google Sites</td>
<td>10,158</td>
</tr>
<tr>
<td>Google</td>
<td>7,594</td>
</tr>
<tr>
<td>YouTube/Other</td>
<td>2,564</td>
</tr>
<tr>
<td>Yahoo! Sites</td>
<td>2,427</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>2,393</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
But what is Search anyhow?
Functional view of a generic technology.

- A Functional model.
  - Shared projects description
  - Shared vocabulary
  - Across Industry, Academia
  - Specific technology positioned within framework
A two pass approach

- **Pass 1: documents**
  - Ingest, index documents, producing Document-meta-data
  - Exploit/augment Document context
  - Build search data-base

- **Pass 2: queries**
  - Prepare query (transform user intent), producing Query-meta-data
  - Exploit/augment query and user context
  - Match
  - Organise results and present them to the user

- **Note**: Real time submission (and processing) of new documents not excluded
Ingest
Document repository

Gather
Document context

Gather
Q/U context

Thematic, Ontology, Type, ...

User Context

Prepare query

Query

Query Refining
Or
Application interaction

Logs
Geo, Device, ...

Interactions

User

D-context

Doc-
ument

Index
Enrich content

Q-meta-data

Match

D-meta-data

Build

Data-base

Build

Present results

Results

CHORUS
AUDIO-VISUAL SEARCH

November 26, 2007
**Key learning: meta-data**

- Meta-data: *any* information contributing to a more synthetic/systematic description of the document
  - Inherited along with document, manually created, automatically generated, not limited to text (feature vectors, …)
  - Meta-data is not restricted to intrinsic document properties: ranking, cluster, theme, …

- Search is about meta-data
  - Matching happens between D-meta-data and Q-meta-data
    - Even « query by example » results in Q-meta-data computations
  - Meta-data must also contribute to query preparation, and helps in results presentation
A definition of search

- Search is about making best use of available meta-data to provide the user with useful information in spite of the fact that his request is possibly poorly formulated and typically unanticipated.
- -> keep the “user in the loop”, maximize his efficiency
**Functional description**

- **Document repository**
- **Ingest**
- **Document context**
  - Gather D-context
  - Geo, Thematic, Language, ...
- **Document**
  - Index Enrich context
  - Geo, Device, ...
  - Interactions
- **Query**
- **Query Refining**
- **Application interaction**
- **User**
  - Prepare query
  - Q-meta-data
  - Match
  - Data-base
  - Build
  - Present results
  - Results
- ** Logs**
Information search: a generic need embedded in a wide range of application areas

- Chorus has grouped information access activities in 6 generic profiles:

<table>
<thead>
<tr>
<th>Category</th>
<th>Unorganized</th>
<th>Semi-organized</th>
<th>Semi-Organized</th>
<th>Organized</th>
<th>Semi-organized</th>
<th>Organized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web Search</td>
<td>Public</td>
<td>Private</td>
<td>Private</td>
<td>Public</td>
<td>Private</td>
<td>Private</td>
</tr>
<tr>
<td>Personalized TV</td>
<td></td>
<td>Unrestricted</td>
<td>Restricted</td>
<td>Unrestricted</td>
<td>Unrestricted</td>
<td>Restricted</td>
</tr>
<tr>
<td>Enterprise Search</td>
<td></td>
<td>Unrestricted</td>
<td>Restricted</td>
<td>Unrestricted</td>
<td>Unrestricted</td>
<td>Restricted</td>
</tr>
<tr>
<td>Public Archives &amp; Digital Asset Mgt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personal Archive Search</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monitoring, Detection &amp; Alert</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*) These categories are neither meant to be exhaustive, nor are they entirely distinct. Several areas may converge on the short term and new applications may emerge.
Information access activities details

WebSearch

- The most visible and successful information access application today. Highly concentrated by a handful of worldwide operators.

Vision

- The rush towards content based multimedia search.
- Increase of share of specialized services for targeted at socio-cultural groups.
- « Disappearing search »: Search will become a back-end facility embedded in front-end user services.

Challenges

- Social threads
  - Increasing awareness of the economic value of activity.
  - User privacy
- System threads
  - Failures
  - Scalability
- Disruptive innovations from other business areas.
Information access activities details

Personalized TV

- Current IPTV services basically offer traditional broadcasting TV and blockbuster movies (via Vod) over ADSL lines.
- A lot of activity among suppliers but a slow adoption curve.
  - A lack of compelling value proposition? No sharing, no annotation, no chat…

Vision

- The TV entertainment component will be integrated with computer and communication environment
  - User will access and share information about content from producers, user annotation and third parties.
- Content recommendation based on user preferences and interactivity appears as key service differentiators.
- Ease of use, and the capacity to share the viewing experience will be key to success.

Challenges

- Ease of use.
- Transitioning of business model from current mass advertising based to a mix of subscription and a targeted advertising.
- Availability of relevant meta-data including time coded tags for efficient information aggregation and navigation.
Information access activities details

Enterprise Search

• While Enterprise Search addresses smaller (but more varied) datasets and lower usage numbers than Websearch and Entertainment services, the corporate users require a very high level of quality.

Vision

• Trend towards take over of Internet techniques on legacy applications will accelerate.
• At work, the corporate users require the same tools and equivalent convenience, with “professional” level performance.
• Search will become an integral part of the Enterprise Information Management toolset,

Challenges

• Transcending the gap between everyday and professional usage.
• Decrease cost of ownership through open standards and architectural choices.
Information access activities details

Public Archives

- As the trend for digitalisation of historical archives accelerates, search is expected to become a major access mechanism.

Vision

- Sooner or later the full historical archives will be on-line.
- Unique value by librarians and manual, editorial value add will remain an essential part of the archives.

Challenges

- Financing access to “cultural heritage”.
- Combining the value add of librarians with automatic indexing and crowdsourcing
  - Technology should “increase”, not “replace”, the added value of librarians and curators.
Information access activities details

**Personal Search**

- A transposition on the personal PC of Web search activities. Few users, few “faces” to recognize, manageable size corpus

**Vision**

- Desktop search remains a “freemium” companion product complementing Web and Enterprise products
- On top of current text-based search, future products will embed image, music, and video classification
  - Rich potential for organizing family picture “shoeboxes” and MP3 archives

**Challenges**

- Business model. Personal search seems in hands of web search majors.
- Image indexing (face/person recognition)
Information access activities details
Monitoring, detection and alerts

• Applications monitoring deviations and or inflexions in observable event streams
  – Typical usage are: weak signal detection, trend analysis, security alerts, IP infringement, competitive survey, ….
• Particularly crucial for political, judiciary and military authorities.

Vision

• Generalization of this domain to consumer applications appears rather far reaching in time
  – Nevertheless some concrete results have been reached recently (e.g.: influenza pandemic detection by Google).
• With the generalization of Web and Internet, the number of available datastreams will explode
  – Regulatory data, Sensors, Social networks, Etc…

Challenges

• Combination of heterogeneous events.
• Lack of standards for modeling situations, events and knowledge representation.
• Quality, confidence and robustness versus attacks.
• Real-time performance.
Search is not the primary innovation drive, but becomes necessary as data volumes grow
  - Search will be everywhere, explicit or implicit, embedded into wider applications

In Future « content-centric » Internet only searchable content will survive
  - Content without meta-data will remain marginal
  - Automatic annotation/structuring and social tagging tools will be essential to extend reach of both users and producers.

Besides researcher genius, money and time, interoperability, cooperation and scalability appear as important success criteria
  - And with mass adoption, improved privacy and trustfulness will become distinctive features.