The Electronic Version of Siku Quanshu

Digital Heritage Publishing Ltd.
Digital Heritage Publishing Ltd

- **Publish Chinese electronic databases**
  - Electronic version of Siku Quanshu
  - General Record of Chinese Culture

- **Develop technology**
  - Chinese computing
  - Digitize Chinese works

- **Collaboration & Consulting services**

- **360 professionals in Hong Kong & Beijing**
  - Management
  - Sales & Marketing
  - Technical
  - Editorial
  - Production
◆ Siku Quanshu
◆ The Electronic Version of Siku Quanshu
◆ Technological Breakthrough
◆ Opportunities
Siku Quanshu
Siku Quanshu

- The Complete Library in Four Branches
  - the 4 classifications of books in ancient China
    - “Jing”  - Classics
    - ”Shi”  - History
    - “Zi”  - Philosophy
    - “Ji”  - Literature
  - largest encyclopedic collection
  - important works in 5000 years Chinese culture
Siku Quanshu

- compiled under the decree of Qing Emperor Qianlong in 1772
  - 3400+ book titles
  - 360+ scholars and 4000 copyists to transcribe
  - 18 years (10 years to complete the first set)
  - 7 sets copied only 3 sets preserved

- 4.7 million pages
- 700 million Chinese characters
The Electronic Version of Siku Quanshu
The Electronic Version of Siku Quanshu

- Comprehensive Searching methods
  - Full Text Search
    - 700 million Characters
    - Boolean search
  - Original classification of Siku Quanshu
  - Book titles
  - Authors
  - Headings (>1.8 million)
The Electronic Version of Siku Quanshu

- Tools
  - Research data management
  - Annotations / Remarks / Notes
  - Editing
  - Viewing
  - Printing
The Electronic Version of Siku Quanshu

- Electronic reference materials
  - Dictionary of Book Abstracts in Siku Quanshu
  - Dictionary of Ancient Chinese Words (with pronunciation)
Technological Breakthrough
Technological Breakthrough

- First Complete Application develop on UNICODE Technology
- Complete Chinese Character Set
- Workflow for digitize Chinese works
  - Innovation of Data Processing
- Cross Platform Technology
Chinese Codes

- Existing Chinese Coding Limitations
  - GB2312 (Simplified Chinese)
    - 6,763 characters
  - Big5 (Traditional Chinese)
    - 13,053 characters
  - Unicode (ISO 10646)
    - CJK + Ext.A (GB & Big5 included)
    - 27,848 characters

- Siku Extended Chinese Character Set
  - 32,000+ characters
  - Unicode based
  - utilize EUDC
Technology

- Siku Extended Chinese Character Set
  - complete & representative

  E.g.  **Guo symbols**

  Ancient Chinese music symbols
Innovation of Data Processing

- Why? OCR (Optical Character Recognition)
  - Project Resources
    - with OCR
      - ~ 800 man-year
    - Traditional Manual Input method
      - >2,000 man-year
  - Strategy
    - 80/20 approach
Innovation of Data Processing

- **Pre-OCR**
  - High Speed Scanning of 2.35 million images
  - Online Image Quality Control system
  - Character segmentation

- **OCR**
  - Data Input for handwritten Chinese

- **Post OCR**
  - Online proof-reading/editing system

- **Format preserved**
Average 22 words per second  (PII 266)
Post-OCR

- Proof-reading
  - by Sequential order
  - by Word
  - by referring original text for contextual clue
Post-OCR

- Page-to-Page Comparison
- Line-to-Line Comparison
OCR & Post-OCR

- OCR Accuracy: 89-91%
- Manual correction
  - select 2nd choice (as suggested by OCR)
    - 1-2%
  - select from among 10 choices
    - 5-6%
- Manual Input
  - 1-2%
Technological Breakthrough

- Format (layout) preserved
  - Image behind Text

- Advanced Character Attribute Setting

- Cross platform technology

- Unicode-based search engine
Technology

- Product Globalization
- Using Single Data/Single Binary Concept
  - Based on Unicode Technology that the same program and data can run on different platforms
  - Unicoding of Data -> SD
  - Unicoding of Programming -> SB
  - Unicoding of UI -> SD
## Cross Platform Performance

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Win 95</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B-</td>
<td>B-</td>
</tr>
<tr>
<td>Win 98</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A-</td>
<td>A-</td>
</tr>
<tr>
<td>Win NT</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>

(The test is using the same program and data runs on different platform)
Technology

- **Unicode Based Search Engine**
  - Handling 30,000+ Character Set
  - Multi-coding Support - Process both Traditional & Simplified Characters in one order
  - Advanced Character Attribute Setting
  - Collected over 200,000+ entries in segmentation database
## Technology

### Search example - related words

<table>
<thead>
<tr>
<th>Phrase to Search without related words</th>
<th>Search Related Words</th>
<th>No. of matches without related words</th>
<th>Related Words</th>
<th>Time (ms) without related words</th>
<th>Related Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>荆軏</td>
<td>荆軏 荆軏</td>
<td>1</td>
<td>42</td>
<td>130</td>
<td>316</td>
</tr>
<tr>
<td>養廉</td>
<td>養廉 羊 檳 養 親</td>
<td>20</td>
<td>45</td>
<td>250</td>
<td>290</td>
</tr>
</tbody>
</table>

from 104 million characters (PII233 128M RAM WinNT 4.0)
Opportunities
Technology Opportunity

- Application Development
  - Second Generation Chinese Search Engine
  - Digital libraries
  - Electronic Professional Database
  - Global market opportunities
ITVentures Ltd.

- Digital Heritage Publishing Ltd.
- Chinese Books Cyberstore Limited
- Cable & Wireless HKT SchoolTeam (Asia) Limited
- Magically, Inc.
- Digital Vision Educational Publishing Company
- Digital Vision Multimedia Company Limited
- Digital Creation Company Limited
The Electronic Version of Siku Quanshu

Web-site:
www.sikuquanshu.com
www.skqs.com
www.dheritage.com