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

迪志文化出版
DIGITAL HERITAGE

- Siku Quanshu
- The Electronic Version of Siku Quanshu
- Technological Breakthrough
- Opportunities

迪志文化出版
DIGITAL HERITAGE

Siku Quanshu

迪志文化出版
DIGITAL HERITAGE

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

迪志文化出版
DIGITAL HERITAGE

Siku Quanshu

- compiled under the decree of

  Qing Emperor Qianlong in1772

  - 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

迪志文化出版
DIGITAL HERITAGE

The Electronic Version of Siku Quanshu

迪志文化出版
DIGITAL HERITAGE

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)

迪志文化出版
DIGITAL HERITAGE

The Electronic Version of Siku Quanshu

- Tools
  - Research data management
  - Annotations / Remarks / Notes
  - Editing
  - Viewing
  - Printing

迪志文化出版
DIGITAL HERITAGE

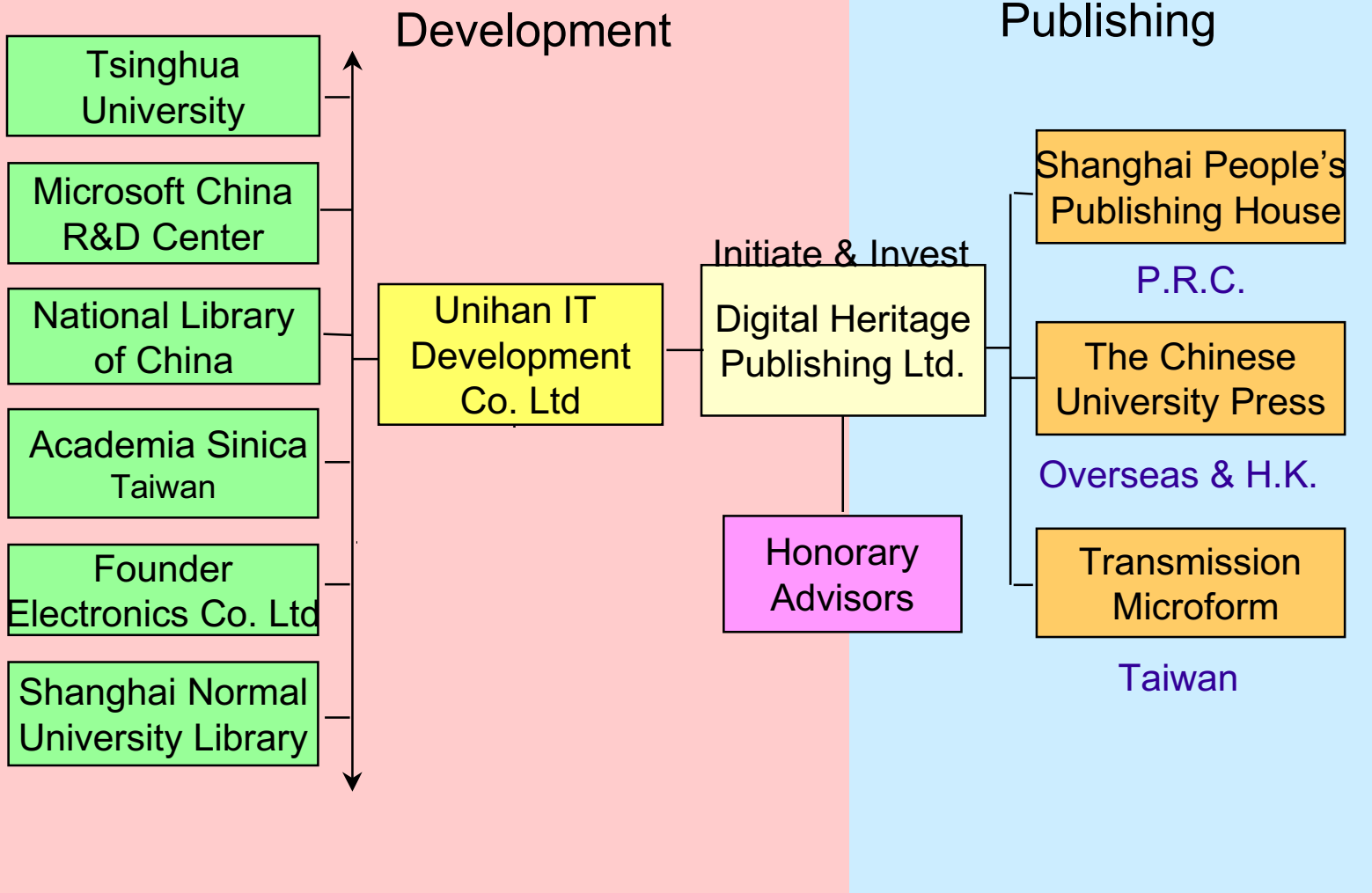
The Electronic Version of Siku Quanshu

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

迪志文化出版
DIGITAL HERITAGE

The Electronic Version of Siku Quanshu

Project Structure chart



Technological Breakthrough

迪志文化出版
DIGITAL HERITAGE

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

迪志文化出版
DIGITAL HERITAGE

Chinese Codes

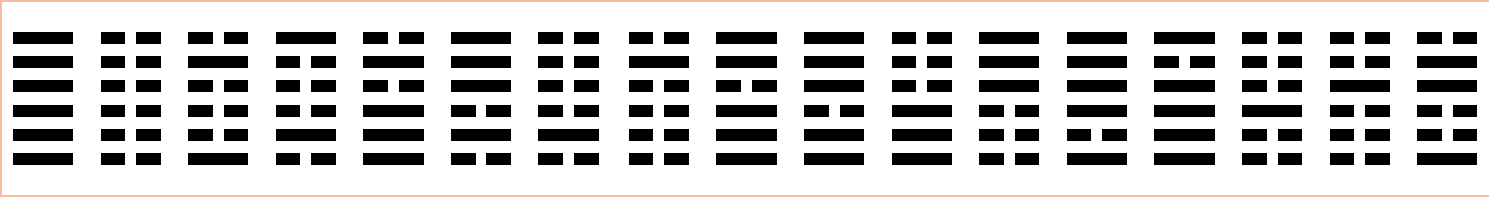
- Existing Chinese Coding Limitations
  - GB2312 (Simplified Chinese)
    - **6,763** characters
  - Big5 (Traditional Chinese)
    - **13,053** characters
  - Unicode (ISO 10646)
    - CJK + Ext.A (GB & Big5 included)
    - **27,848** characters
- *Siku Extended Chinese Character Set*
  - **32,000+** characters
  - Unicode based
  - utilize EUDC

迪志文化出版
DIGITAL HERITAGE

Technology

- Siku Extended Chinese Character Set
  - complete & representative

E.g. Guo symbols



Ancient Chinese music symbols

迪志文化出版
DIGITAL HERITAGE

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

迪志文化出版
DIGITAL HERITAGE

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

迪志文化出版
DIGITAL HERITAGE

Pre-OCR

- Character segmentation

OCR

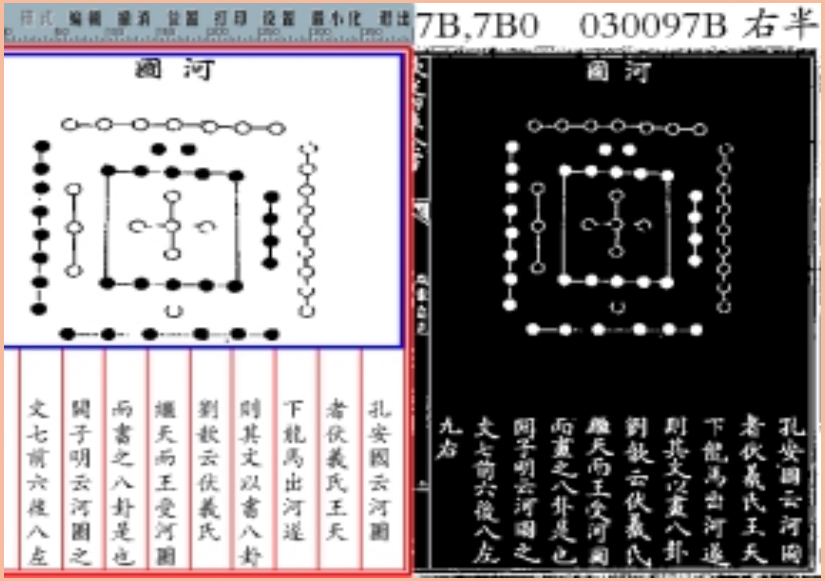
Average 22 words per second (PII 266)

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

迪志文化出版
DIGITAL HERITAGE

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

迪志文化出版
DIGITAL HERITAGE

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

迪志文化出版
DIGITAL HERITAGE

Cross Platform Performance

	Simplified Chinese Windows	Traditional Chinese Windows	Japanese Windows	Korean Windows	English Windows
Win 95	B	B	B	B-	B-
Win 98	A	A	A	A-	A-
Win NT	A	A	A	A	A

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

迪志文化出版
DIGITAL HERITAGE

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

迪志文化出版
DIGITAL HERITAGE

Technology

■ Search example - related words

Phrase to Search		No. of matches		Time (ms)	
without related words	Related Words	without related words	Related Words	without related words	Related Words
荆 軻	荆 軻 荆 轲	1	42	130	316
養 廉	養 廉 羣 槏 养 廉 養 覝	20	45	250	290

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

Opportunities

Technology Opportunity

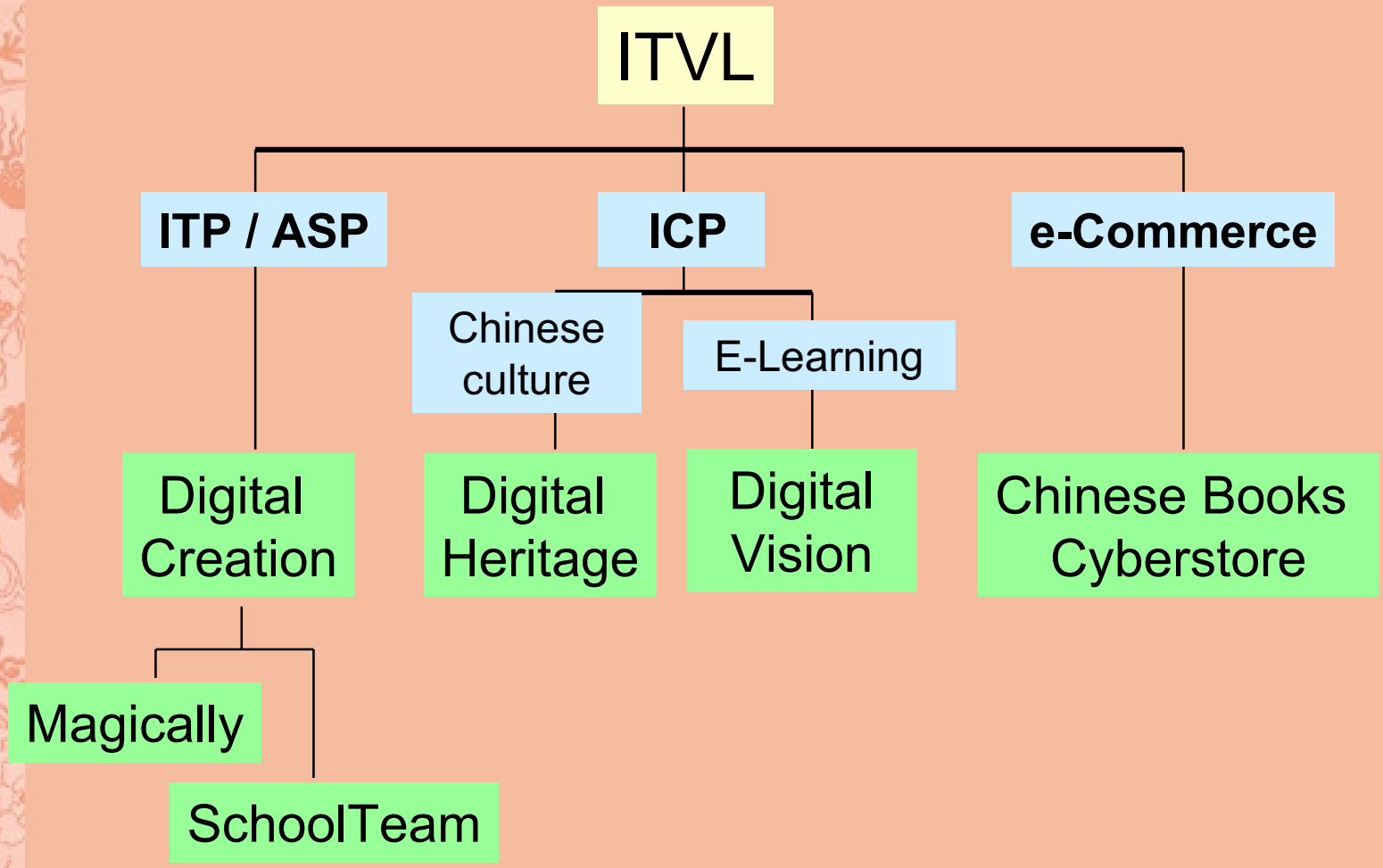
- **Application Development**
  - Second Generation Chinese Search Engine
  - Digital libraries
  - Electronic Professional Database
  - Global market opportunities

迪志文化出版
DIGITAL HERITAGE

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

ITVentures Ltd.



The Electronic Version of Siku Quanshu

Web-site:

www.sikuquanshu.com

www.skqs.com

www.dheritage.com

迪志文化出版
DIGITAL HERITAGE