

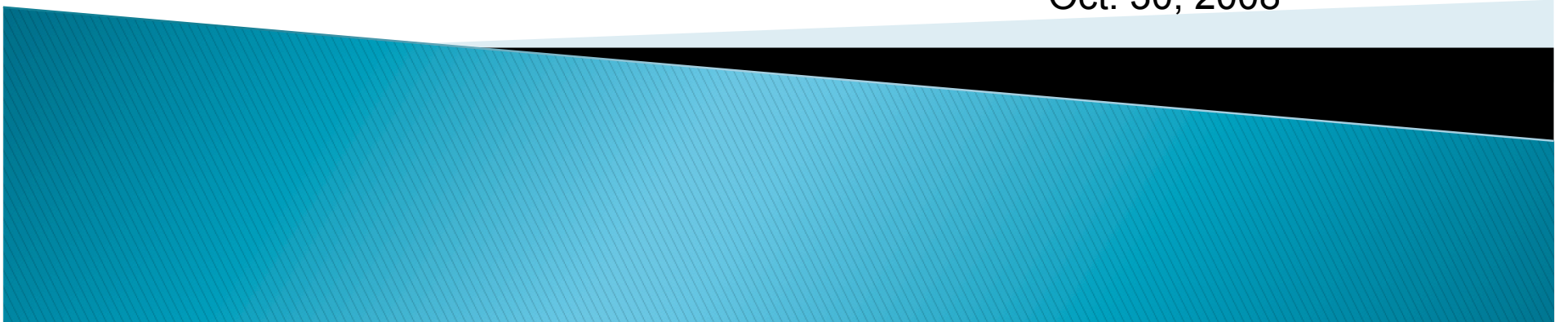
UC San Diego Google Mass Digitization Project

Jim Cheng

University of California San Diego

2008 PRDLA Meeting, Singapore

Oct. 30, 2008

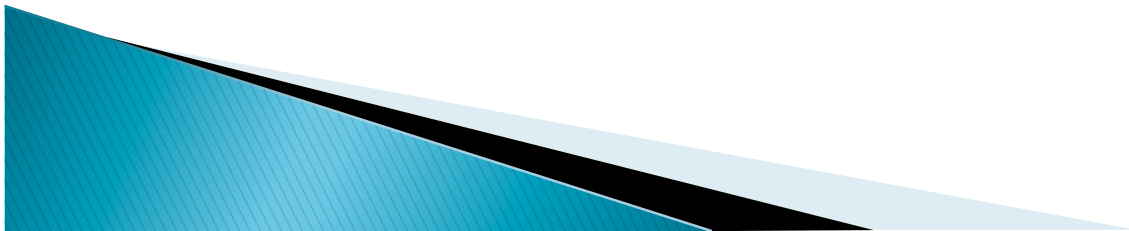


Historical Background

The UC Google Project targeted the IR/PS (International Relations & Pacific Studies) and EA (East Asia) Collections at UCSD as initial objects of its pilot endeavor to scan from campus collections. It started in April 2008.

1. Google Partnership:

- ❑ The initial Google 5: Oxford, Harvard, U. Mich., NYPL, and Stanford in 2004 <http://books.google.com>.
- ❑ UC became a partner in 2006 for digitizing 2.5 million volumes of its collections during a 6-year period.
- ❑ Today, the Google 25 (including CIC Consortia) are involved with the Google Book Project with more than 2 million digitized volumes, at a scanning rate of 3,000 volumes/day.



Historical Background

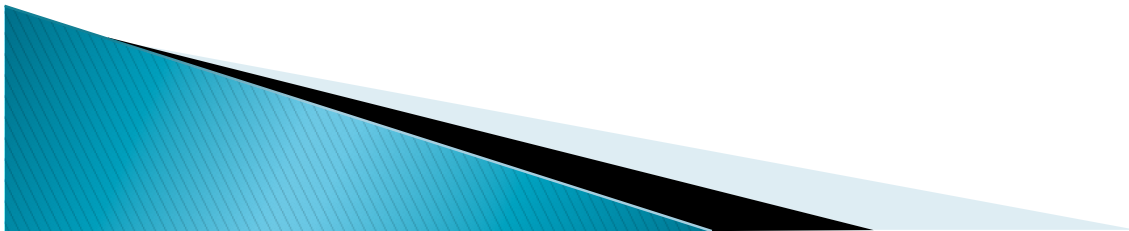
2. Involved East Asia Collections:

- Currently, there are 4 EA Collections in North America (Michigan, Harvard, Stanford, and UCSD) involved with the Google Book Search Mass Digitization Project.

Different models:

- Restricted digitization: Harvard (before 1909),
- Complete digitization (except non-standard and publisher opt-out materials): Michigan, Stanford, and UCSD.

Total: More than 1.5 million volumes in CJK (Chinese, Japanese, and Korean) and other languages.



Historical Background

3. Other Major Open Sources:

- ❑ Microsoft Live Search Books (2006–2008) 750,000 books and 80 million journal articles. The web site shut down on 5/27/08, but the content is accessible through Internet Archive. Only out-of-copyright books and no CJK materials are included.
- ❑ OCA (Open Content Alliance) / Yahoo / Internet Archive (2005–) <http://www.archive.org/details/opencontentalliance>: More than 100 libraries with 8 scanning centers in 3 countries are involved with more than 200,000 books (including the fold-outs), are digitized. Digitized materials include microfilm, out-of-copyright and out-of-print books, and documentary films. Yahoo is responsible for indexing these materials. So far, there are no CJK materials. All Microsoft Live Search Books will be ported to Internet Archive.
- ❑ US–China Million Book Digital Library Project <http://www.cadal.zju.edu.cn/Index.action>:
 - Classical Texts: 古籍 (146,669 v.),
 - Books published between 1911–1949: 民国图书 (181,977 v.),
 - Journals published between 1911–1949: 民国期刊 (3802 titles),
 - English books published before 1923 (130,525 volumes), paintings (3,427).(Need to download the viewing software: djvu.)

UCSD Google Project

1. Major features:

- Based on UCSD's recommendation, Google conducted its collection analysis and selected the IR/PS and EA Collections as the first mass digitization project for active collections among southern UC campuses, which started in April 2008 and ends in the summer of 2009.

- Established in 1987, the IR/PS Library is the only academic library in the United States to focus primarily on the contemporary economies, politics, and markets of the Pacific Rim. With regional strengths in East Asia, Southeast Asia, and Latin America, the collection features more than 140,000 bound volumes, and some 1,400 active periodical subscriptions in English, Chinese, Japanese, Korean, Spanish, Portuguese, and other languages.

- Established in 1988, the East Asia Collection focuses on collecting modern and contemporary Chinese, Japanese and Korean (CJK) language materials in the humanities and social sciences, including literature, history, sociology, linguistics, and philosophy pertaining to the region. With over 150,000 volumes in print,.....

UCSD Google Project

1. Major features (cont.):

- ❑ Google hosts all digitized titles at Google Book Search <http://books.google.com/> and provides full-text searchable access.
- ❑ UC/CDL receives a copy of each digitized title.
- ❑ UC/CDL (California Digital Library) keeps all copies from Google in the MDID (Mass Digitization Inventory Database) and DPR (Digital Preservation Repository) for access and preservation purposes.
- ❑ UC Libraries have the right to use the UC Libraries digital copy, in whole or in part at University's sole discretion, subject to copyright law, as part of the services offered to University Library patrons.
- ❑ UC Libraries are permitted to distribute no more than 10% of the UC Libraries digital copies to other libraries and educational institutions for non-commercial, research, scholarly, or academic purposes (but not any portion of image coordinates).

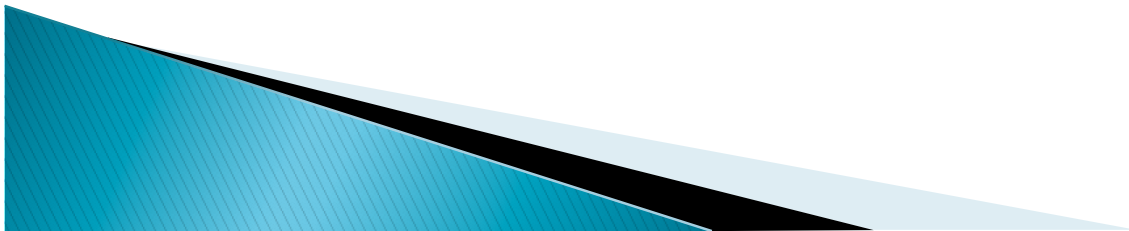


UCSD Google Project

1. Major features (cont.):

- Different views based on copyright and publisher agreements:
 - Full View – items published before 1923 or are public domain materials.
 - Limited View – provided by the Google/Publisher Partnership with up to 75% of the contents in display and searchable mode with purchase and borrowing information.
 - Snippet View – three citations matching the search term, and full-text searchable with purchase and borrowing information.
 - No Preview – only metadata available with purchase and borrowing information.

Sample keyword search: 孔子 for the above viewing models



UCSD Google Project

1. Major features (cont.):

- Integrated access with other products:
 - The Google Book Search results display with other resources, such as Google Map, Amazon, and OCLC WorldCat.

Sample keyword search: China earthquake/full view

**Results :Journal of the Statistical Society of London, v. 41 /1878
(Oxford)**

Problems: No name authority control for geographical names, such as Canton (Guangzhou), and Hankow (Hankou)

Sample keyword search: An Annotated Bibliography for Chinese Film Studies

Problems: Jinhua (confused with the personal name Dai Jinhua and the city name Jinhua).



Upper part of the result page:

The screenshot shows a Microsoft Internet Explorer browser window displaying a Google Book Search result. The address bar shows the URL: <http://books.google.com/books?id=QawEAAAQAAJ>. The page title is "Journal of the Statistical Society of London" by Statistical Society (Great Britain). The page is divided into several sections:

- About this book:** Includes a book cover image, author information (Statistical Society, Great Britain), publication year (1878), original source (Oxford University), and digitization date (Oct 31, 2006). It also features buttons for "Read this book" and "Download PDF", and links to "Add to my library" and "Write review".
- Buy this book:** A list of links to purchase the book from various retailers: Abebooks, Alibris, Amazon, Barnes&Noble.com, and Google Product Search.
- Borrow this book:** A link to "Find this book in a library".
- Contents:** A list of page numbers and corresponding titles, such as "VOL XLLYEAR" (page 3), "United Kingdom, income tax, talise" (page 32), "Discussion on Mr Giffens Paper" (page 40), "Diagrams Exhibiting the Positions of the Bank of England" (page 83), "Discussion on Mr Seyds Paper" (page 113), "Discussion on Mr Mundellas Paper" (page 121), and "Adjourned Discussion on Mr Mundellas Pajxr" (page 121).
- Selected pages:** Three thumbnail images of the book's title page, table of contents, and index, with links below them.
- Search in this book:** A search box with a "Search" button.
- Other editions:** A link to "Journal of the Statistical Society of London" by Statistical Society (Great Britain) - Statistics - 1881.

The browser's taskbar at the bottom shows several open applications, including "Inbox - Micr...", "Re: proofre...", "LWSummer...", "Journal of t...", and "Desktop". The system tray on the right shows the time as 11:40 AM and the date as EN.

Bottom part of the result page

Journal of the Statistical Society ... - Google Book Search - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://books.google.com/books?id=QawEAAAQAAJ> Go Links

Google [google book search](#) [Go](#) [Bookmarks](#) [801 blocked](#) [ABC Check](#) [Look for Map](#) [AutoFill](#) [Send to](#) [google](#) [book](#) [search](#) [Settings](#)

by Statistical Society (Great Britain) - [Statistics](#) - 1883
Published by: Charles Knight and Co., 1838-Jan. 1842; John William Parker, Apr. 1842-1860; Edward Stanford, 1861-1886.
[Full view](#) - [About this book](#) - [Add to my library](#)

[show more >](#)

Key terms

[United Kingdom](#), [earthquake](#), [Bank of England](#), [Owens College](#), [kilogrammes](#), [Scotland](#), [Comets](#), [Ireland](#), [hailstorm](#), [political economy](#), [Italy](#), [bullion](#), [Russia](#), [drought](#), [Roma](#), [free trade](#), [famines in India](#), [Germany](#), [Baring Brothers](#), [Austria](#)

Places mentioned in this book

[Manchester](#) - [Page 540](#)
Henry, took their degrees in medicine with honour at Edinburgh, and various sours of **Manchester** bankers and manufacturers have resorted to the Scotch ...
more pages: [577](#)

[Glasgow](#) - [Page 558](#)
more conclusive experiment could have been planned than that which has been unintentionally carried out among the inhabitants of **Glasgow** and Hillhead. ...
more pages: [554](#)

[Madras](#) - [Page 525](#)
Hunter's results, for as the cycle of rainfall at **Madras** coincides, I am informed, with the periodicity of the cyclones in the adjoining Bay of Bengal ...

[more >](#)

[About Google Book Search](#) - [Book Search Blog](#) - [Information for Publishers](#) - [Provide Feedback](#) - [Google Home](#)

©2008 Google

Done Local intranet

start [Inbox - Micr...](#) [Re: proofre...](#) [UWSummer...](#) [Journal of t...](#) Desktop >> EN [Local intranet](#) 11:25 AM

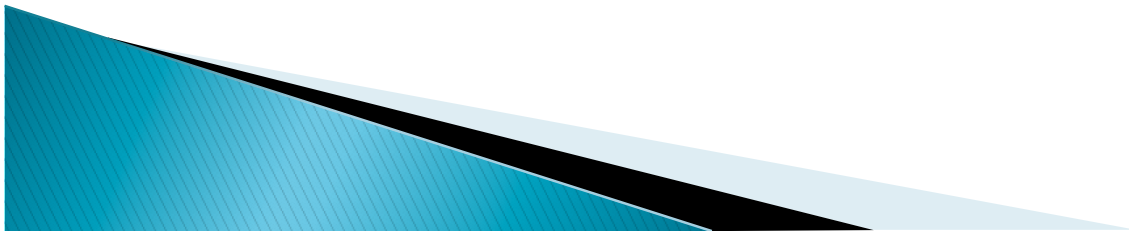
UCSD Google Project

1. Major features (cont.):

- Google Book Search API (Application Programming Interface):
 - UC's initial implementation:
API enables the embedded online access to Google Book Search information from UC's OPAC/Melvyl.
 - UC's further implementation:
An embedded YouTube-style viewer for Google books is implemented that enables the users to view the related Google book information within Melvyl.
 - Also in Roger, the UCSD local catalog.

Sample keyword search: "An Annotated Bibliography for Chinese Film Studies"
on UC OPAC

<http://melvyl-test.cdlib.org:8164/F/>



Google API used in Melvyl (UC OPAC)

Melvyl - Search Results - Microsoft Internet Explorer

Address: http://melvyl.cdlib.org/F/N5SMS253NULXUQ8XICJ5K3AY9KNIGTNRJXX9KQE43TK6QJAT1J9-00508?func=find-b&find_code=WTI&request=annotated+bibliography+for+Chinese+film+studies&adjacent=Y&filter_code_4=WID&

Melvyl® The Catalog of the University of California Libraries

Search results: **1 Item(s)** [Modify Search](#)

Searched: **Title= annotated bibliography for Chinese film studies (Phrase)**
Collection: Entire Collection

[Print / Email](#) [Save](#) [Request](#)
[select all](#) [deselect all](#)


[1 Details/Locations](#)


Author [Cheng, Jim](#)
Title An annotated bibliography for Chinese film studies = Zhongguo dian ying yan jiu shu mu ti yao / Jim Cheng
An annotated bibliography for Chinese film studies = 中国电影研究书目提要 / Jim Cheng
Publisher Hong Kong : Hong Kong University Press, c2004
Format Book
Library UCB UCSD UCD UCLA UCSC UCSB UCR UCI

[Print / Email](#) [Save](#) [Request](#)
[select all](#) [deselect all](#)

Display: [Short](#) [Long](#) [Review](#)
Sort: [Year](#) [Author](#) [Title](#) [Uniform Title](#)
Sorted by:
Sorting and display limited to first 1,000 records

[Previous](#) [Next](#) Item # [Go](#)

 [Google Book Search](#)
[Preview this book online!](#)



CDL
Comments and feedback
Privacy Policy
Melvyl® is an initiative of the California Digital Library
© 2006 The Regents of the University of California

Local intranet

Google API used in Melvyl (UC OPAC)

Melvyl - Full View of Record - Windows Internet Explorer provided by Yahoo!

http://melvyl-test.cdlib.org:8164/F/MI466LRGIHIAYXQLR6CH2VCS9RMX8TCCSDC3V8UC85ENSG9U6-02992?func=full-set&set_number=001296&set_entry=000001&format

File Edit View Favorites Tools Help

Google

DivX

Basic Search Advanced Command Browse Most Recent Search Previous Searches Saved Items

This is a test version of Melvyl. Please do not use the Request, Update, Profile, and My Workspace services. All other functions are available.

Search results: 1 Item(s) [Modify Search](#)

Display: [Full](#) [MARC](#)

[Print / Email](#) [Save](#) [Save Across Sessions](#) [Request](#)

[Previous](#) [Next](#)

Item 1 of 1 Total

[Return to Search Results List](#)

Author [Cheng, Jim](#)

Title An annotated bibliography for Chinese film studies = Zhongguo dian ying yan jiu shu mu ti yao / Jim Cheng

An annotated bibliography for Chinese film studies = 中國電影研究書目提要 / Jim Cheng

Publisher Hong Kong : Hong Kong University Press, c2004

Description xii, 404 p. ; 28 cm

Note Includes bibliographical references and indexes

ISBN 9622097030 (hbk.)

Language English

Subject [Motion pictures -- China -- Bibliography](#)

[Motion picture industry -- China -- Bibliography](#)

[Film criticism -- China -- Bibliography](#)

Format Book

Library [UC Berkeley](#) [UC San Diego](#) [UC Davis](#) [UC Los Angeles](#) [UC Santa Cruz](#) [UC Santa Barbara](#) [UC Riverside](#) [UC Irvine](#) [All](#)

full-screen disable

Search

An Annotated Bibliography for Chinese Film Studies

中國電影研究書目提要

著 Jim Cheng 程傑

Google Book Search Buy this book More about this book

Library	Call Number	Availability	Notes

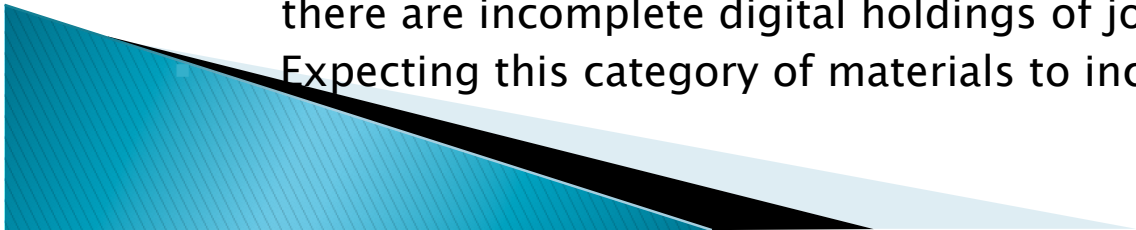
Local intranet 100%

start

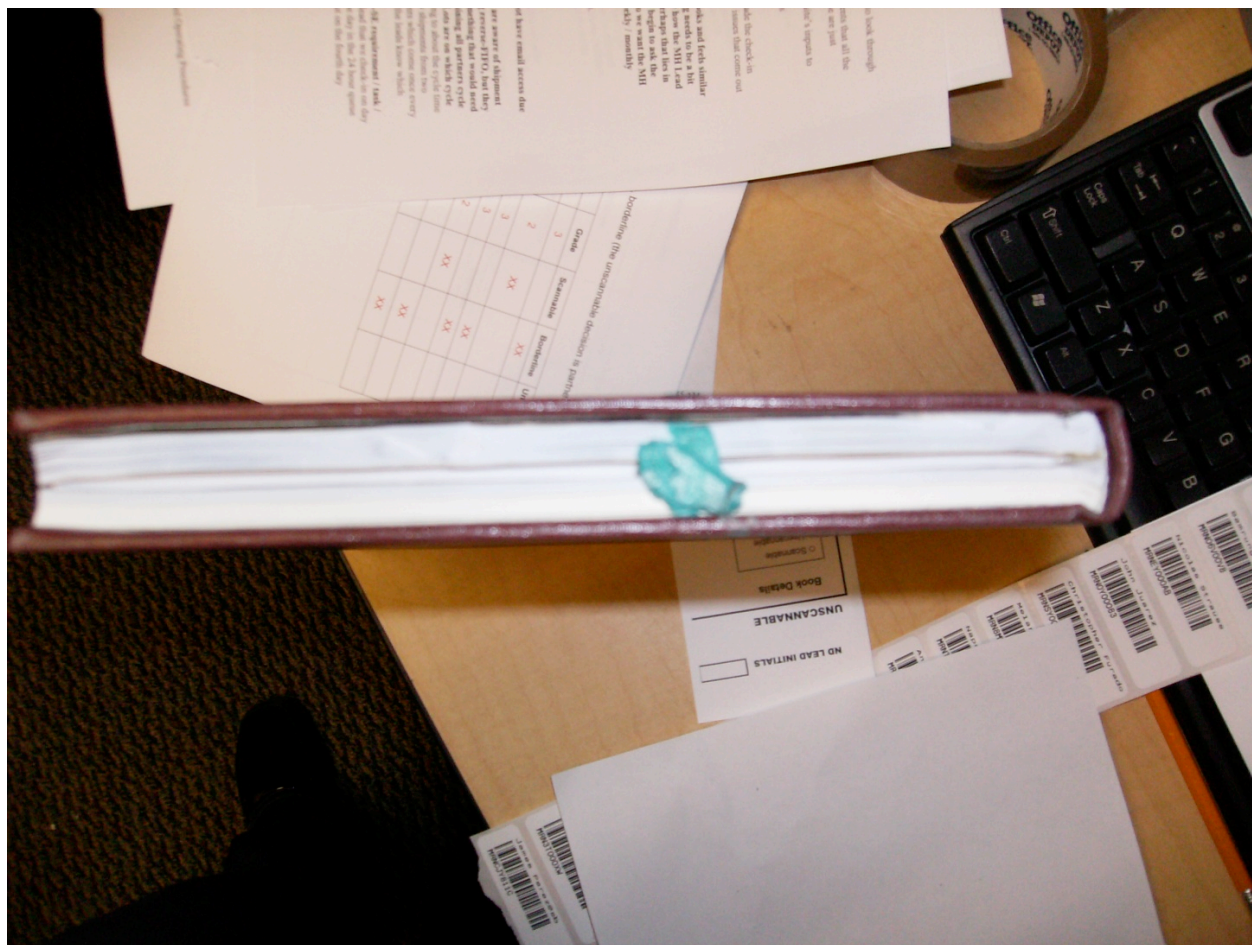
Desktop EN 2:19 PM

UCSD Google Project

2. Major concerns:

- Applying US copyright to the CJK materials.
 - Politically sensitive materials in different nations:
Sample author search: Hitler's works are illegal in Germany and Fa Lun Gong/Li Hongzhi's materials are illegal in China
 - Foldouts are not scanned (2% thru 10/1/08):
 - Materials not matching the Google scanning standard (4% thru 10/1/08):
 - oversize, mini-size, maps, newspapers (including reprints), pictorial and other materials.
 - Publisher opt-out materials (4% thru 10/1/08):
 - The issue had not been initially announced by Google,
 - Specifically, some articles in bound volumes of journals. As a result, there are incomplete digital holdings of journals.
Expecting this category of materials to increase.
- 

“Gummy” book rejected by Google



UCSD Google Project

2. Major concerns:

- The Metadata for non-Roman materials (such as CJK materials) originally sent to Google lacked the vernacular fields (880s).
 - Presently, these materials can only be found by the Romanized terms in the regular metadata search, such as Title, Author, and Publisher, but not by using original language character search. We have now sent complete metadata and expect this to be corrected in the next month or so.

Sample title search: Guo ji jing ji he zuo/International Economic Cooperation/國際經濟合作

- Treating everything as a book:
 - No serials holdings statement,
 - No title change statement, and
 - No logical listing for different issues of the same serial title.
 - No gathering of serials volume issues as normally presented in library catalogs.

Sample title search: Malaya Economic Review (continued by Singapore Economic Review in 1983)



UCSD Google Project

2. Major Concerns (cont):

- Diacritic-sensitive non-English romanized term search:

Sample title search:

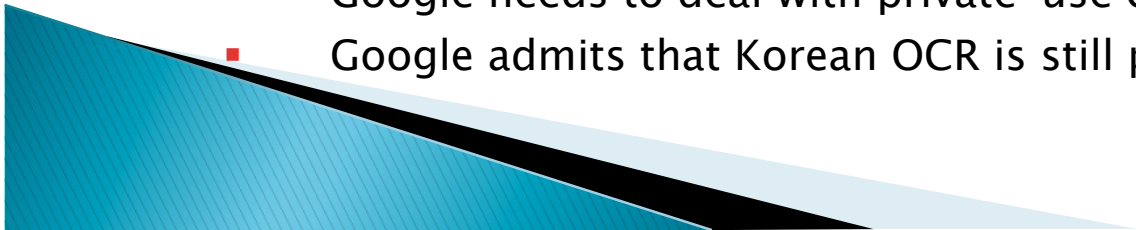
El Pacífico Sur como parte de una potencial comunidad economica del Pacífico (Spanish), Nü xing fan zui shi lu (Chinese), and Kyŏngje nonjip (Korean)

- No combined simplified and traditional Chinese search

Sample title search:

国际经济 / 國際經濟

- Unicode with CJK OCR (without viewers and plug-ins):

- Unicode 5.0 supports some 90,000 Han characters.
 - Example: Digital Heritage Product: SKQS Version 3.0 uses 70,195 public-use and 12,592 private-use characters, and needs to download the character set as plug-ins in order to reach 99% searchable functionality.
 - Google needs to deal with private-use characters.
 - Google admits that Korean OCR is still problematic.
- 

UCSD Google Project

3. Future Impacts:

- ❑ impact as a new evolved model of digital library service,
- ❑ impact on library collection development,
- ❑ impact on library reference services,
- ❑ impact on library preservation,
- ❑ impact on library user behavior and discovery of library resources
- ❑ impact on library OPAC and Web page development.



End

