Shades of Grey

Presented by

Bonnie C. Carroll, President
Information International Associates, Inc.
Oak Ridge, TN

Presented to

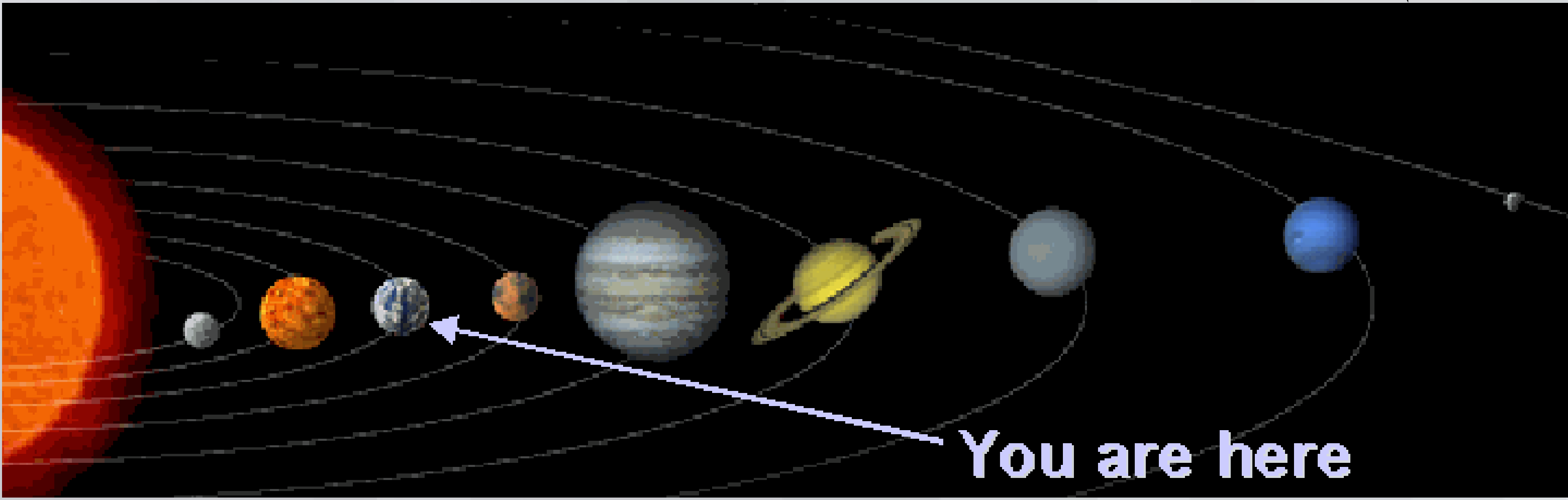
GL8 CONFERENCE
Eighth International Conference on Grey Literature
"Harnessing the Power of Grey"
4-5 December 2006
Lindy C. Boggs International Conference Center
New Orleans, Louisiana USA

Information International Associates, Inc. (IIA)

Nothing new under the sun

…but we live in an expanding universe of grey

Disclaimer



On August 24th, 2006 Pluto's status was officially changed from planet to dwarf planet. For decades children have been taught that there are nine planets in the Solar System. However, with this change, there are now only eight planets.

The Electronic World cannot be dealt with as a linear extension of the print world.

To understand the grey literature of the future, one has to understand how communications and documentation are taking place.

Shades of Grey

- Spectrum of changing characteristics
- A filter or a curtain

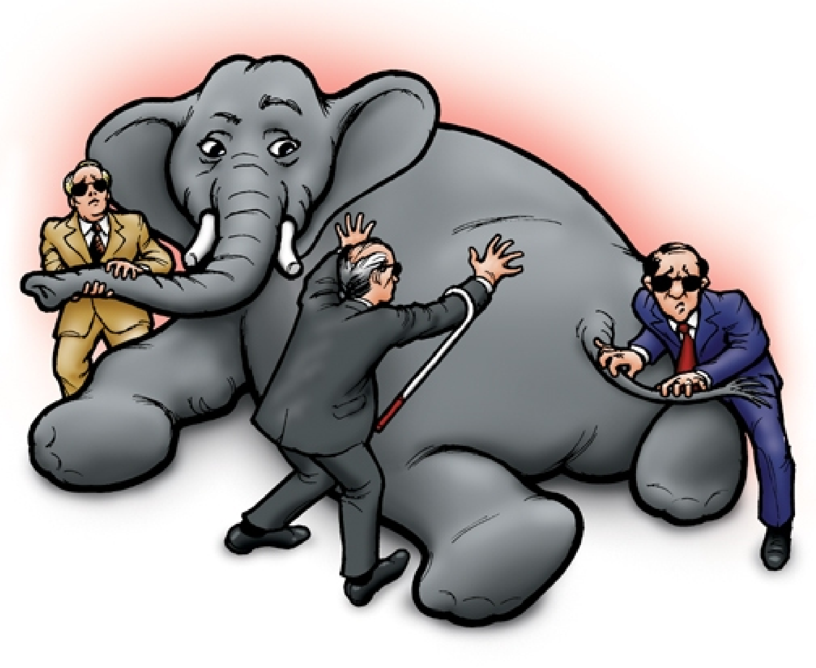
Information International Associates, Inc. (IIA)

Theses of This Talk

- Grey literature in traditional media may still be a problem, but
- The new networked world is causing a shifting paradigm and little is black and white
- Definitions are dynamic
- Challenges are exciting

Background Observation: Grey Literature is in the eye of the beholder

- By profession
- By subject matter
- National intelligence
- Business intelligence
- Academia
- Librarians
- Law
- Science
- Public Health – NY Academy of Medicine Grey Lit Report
- Every discipline (overall the literature of agronomy is 40% grey lit) (Gerry McKierman, 2003)
- Every Community



Rip Van Winkle

opening ones eyes 20 years later

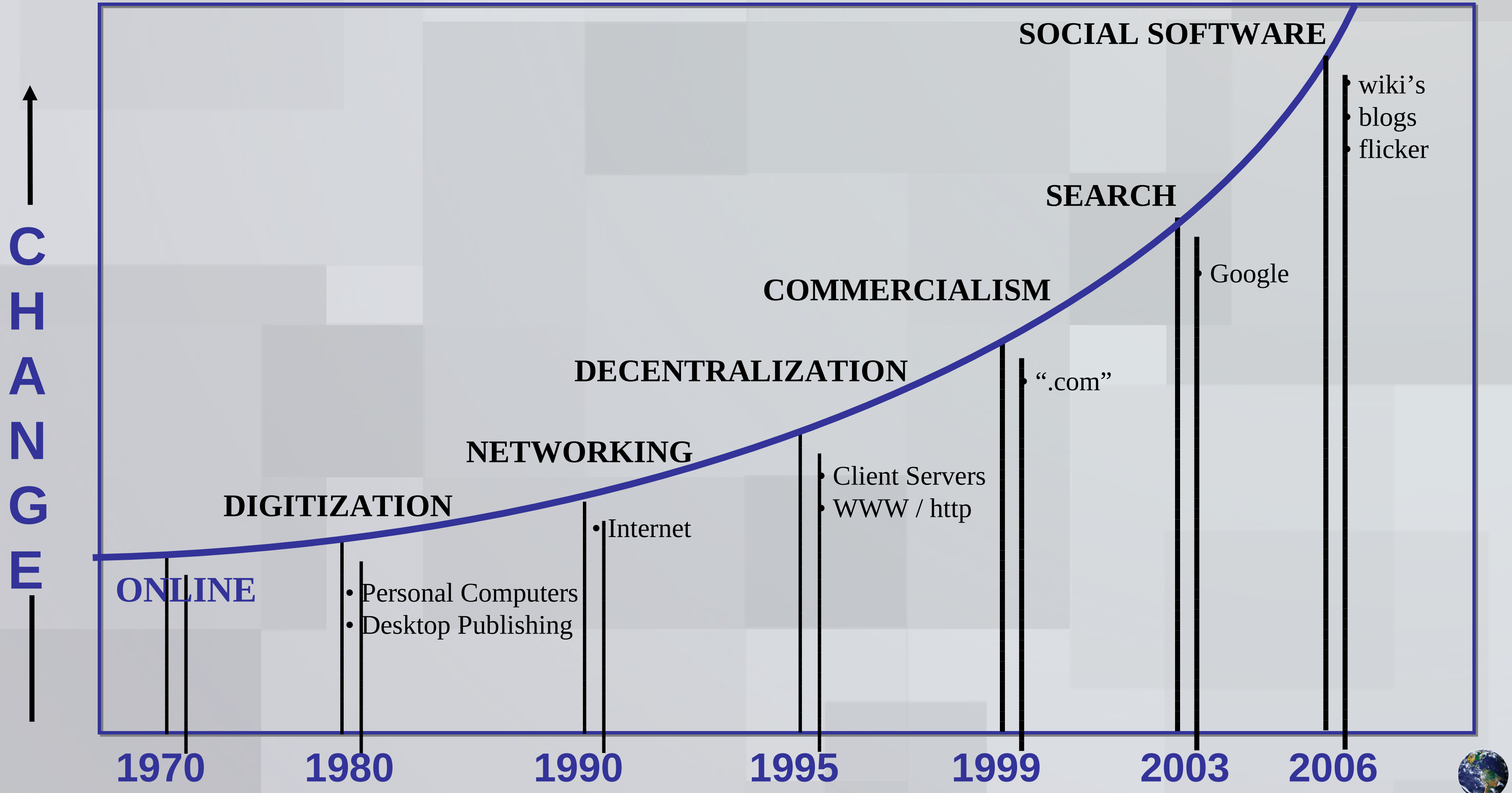
- Technological Change
- Volume of Content

Rapidly advancing technologies have opened new opportunities

Five years ago, right after the start of the personal computer revolution, industry experts observed that if the automobile business had developed like the computer business, a Rolls Royce would cost $2.75 and go 3 million miles on a gallon of gasoline.”

-- *Fortune* Magazine
August 1, 1998, p. 4

TECHNOLOGIES OVER TIME



CHANGE

ONLINE

DIGITIZATION

- Personal Computers
- Desktop Publishing

NETWORKING

- Internet

DECENTRALIZATION

- Client Servers
- WWW / http

COMMERCIALISM

- ".com"

SEARCH

- Google

SOCIAL SOFTWARE

- wiki's
- blogs
- flicker

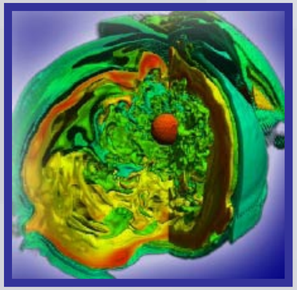
Volume: In 1997 the Library of Congress has 3 petabytes ($3x10^{15}$)

Calculation

- 20 TB-20M books x IMG each
- 13 TB-13M Photos, compressed to 1MB JPG each
- 200TB – 4M Maps scanned
- 500TB – 5M Movies, 1GB each
- 2,000 TB – 3.5M Sound Recordings, 1 audio CD each

http://www.lesk.com/mlesk/ksg97/ksg.html

About Volume: How much Data is there?



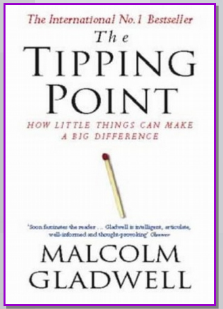
TeraScale Supernova Initiative = 5 terabytes per day

iPod Shuffle (up to 120 songs) = 512 MegaBytes

Printed materials in the Library of Congress = 10 TeraBytes
OSTI = 3 TeraBytes

Kilo	10^3
Mega	10^6
Giga	10^9
Tera	10^{12}
Peta	10^{15}
Exa	10^{18}
Yatta	10^{24}

1 small novel = 1 MegaByte



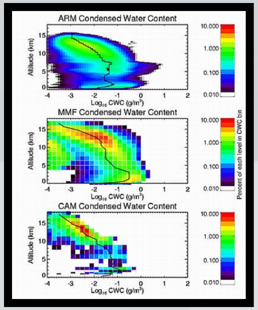
The International No.1 Bestseller

The TIPPING POINT

HOW LITTLE THINGS CAN MAKE A BIG DIFFERENCE

MALCOLM GLADWELL

Atmospheric Radiation Measurement Program (ARM) Data Archive = 41 terabytes



1 Low Resolution Photo = 100 KiloBytes

All worldwide information in one year = 5 ExaBytes

Slide compliments of Walt Warnick and based on Fran Berman, UCSD

Volume: How big is five exabytes?

- 5 exabytes of new information in 2002
  - Print, film, magnetic and optical storage media
  - 92% on magnetic media
- If digitized, the 19M books and other print collections in the Library of Congress would contain about 10TB of information
- 5 exabytes is equivalent in size to ½M new libraries the size of the LC print collections.

file://///Herald/www/research/projects/how-much-info-2003/execsum.htm (1of14) [10/30/2003 4:11:03 PM]

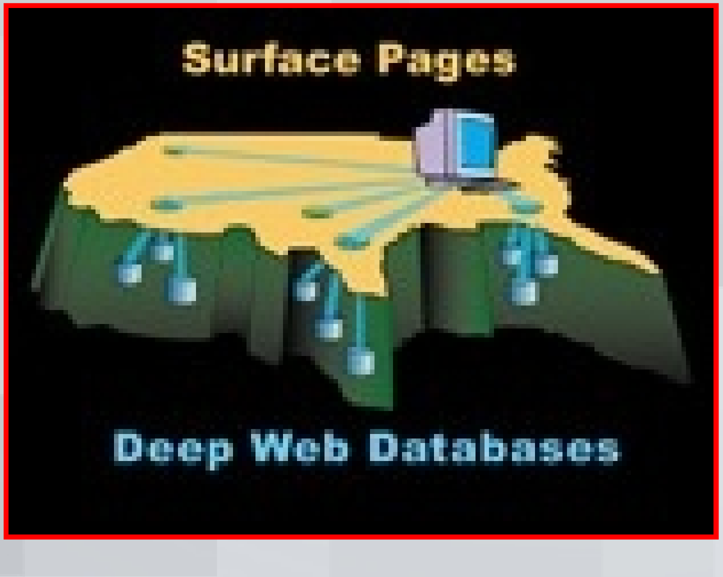
The www: Fastest growing new publishing media of all time and medium of first resort

- Surface Web
  - Static, publicly available
- Deep Web
  - Dynamic, database driven
- Public Web
  - meant for dissemination
- But the **Key To Access Is Search**
  - Depends on proprietary algorithms
  - Limited coverage of deep web

In 2003, UC Berkeley's School of Information Management and Systems estimated with 95 percent confidence:

surface web: **167 TB** (10^{12})

~~*deep Web*: between~~
66,800 TB and **91,850 TB**
>1%

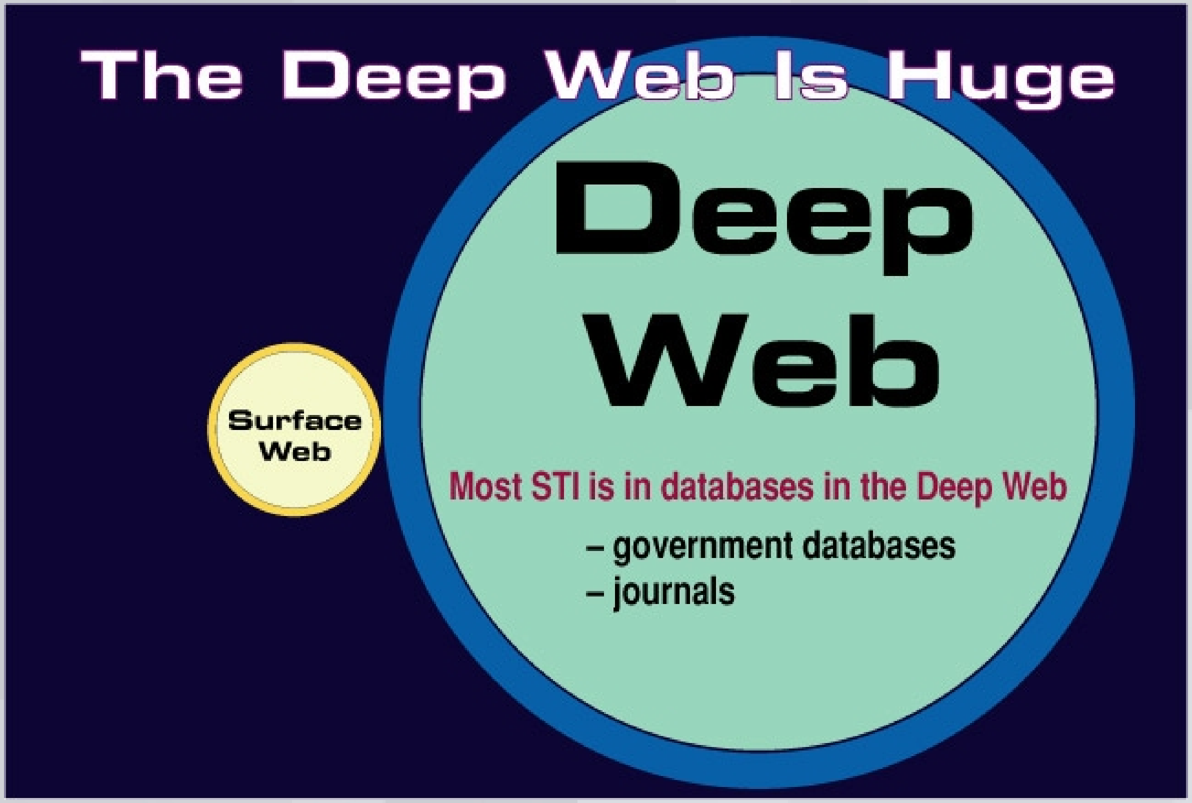


and growing!

Slide compliments of Walt Warnick

Getting to the R&D Results

- The Surface Web is accessible to popular search engines such as Google.
- But less than 1% of government R&D results are currently accessible to crawlers.



The Deep Web Is Huge

Surface Web

Deep Web

Most STI is in databases in the Deep Web

– government databases
– journals

Slide compliments of Walt Warnick

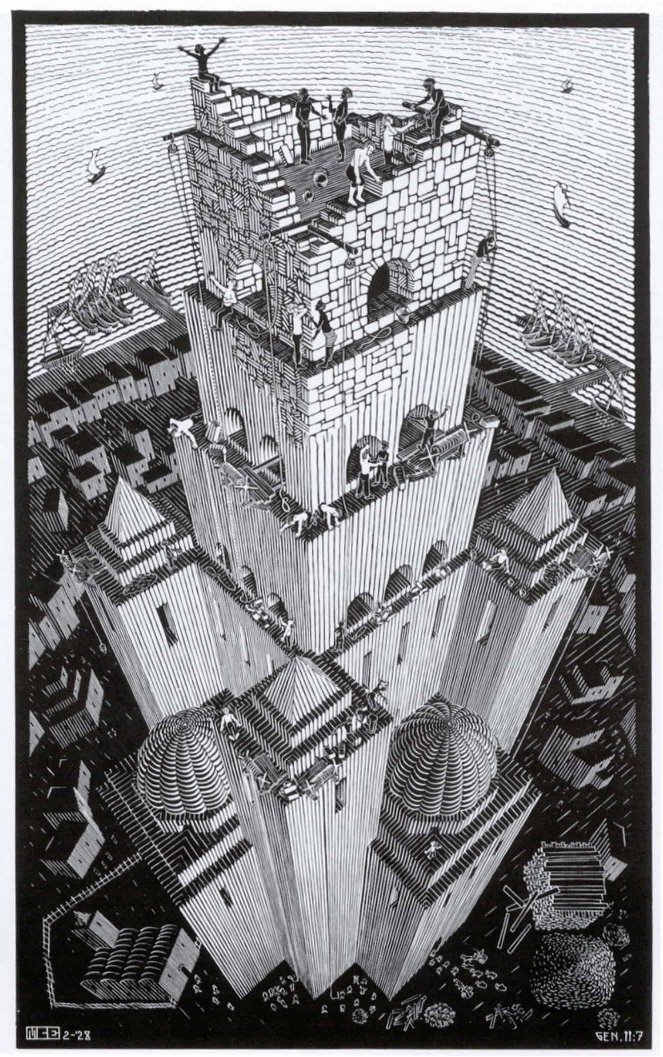
Surface Pages
Deep Web Databases

- Deep Web will grow
- Huge data collections are useful only if patron knows where to find them
  - What shade of grey?
- Professional librarians go from Dialog to Deep Web
  - But, increasing end user information literacy redefines the shade of grey

Slide compliments of Walt Warnick

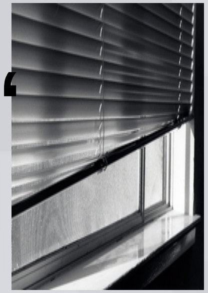
Definitions: A Tower of Babel

- “Grey” – most of talk
- “Literature” a container word
- Language barrier – languages of the internet
- Changing definitions of publisher



Traditional Concepts of Grey Literature

- Definitions
- Unique Characteristics



“Greynet Definition of Grey Literature.” Dominic Farace 1997

“That which is produced on all levels of government, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishers.”

L

reaffirmed GL6 in NY 2004

Publications are not controlled by commercial venders & publishing is not the primary business activity of the producing bodies.

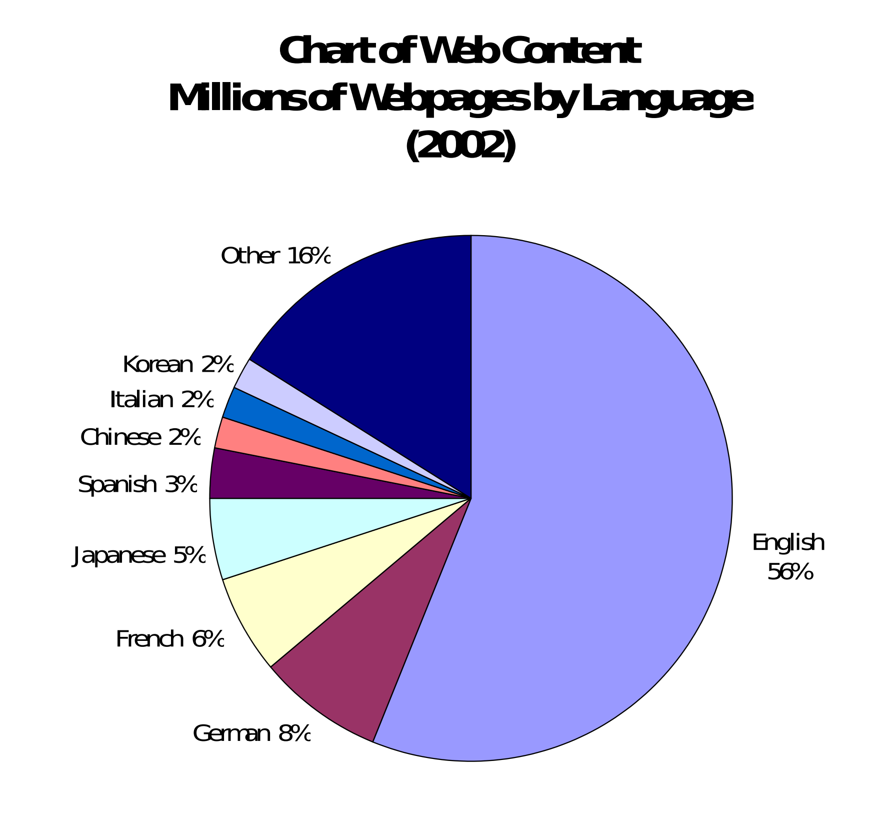
N

Access to Digital Information (PADI)

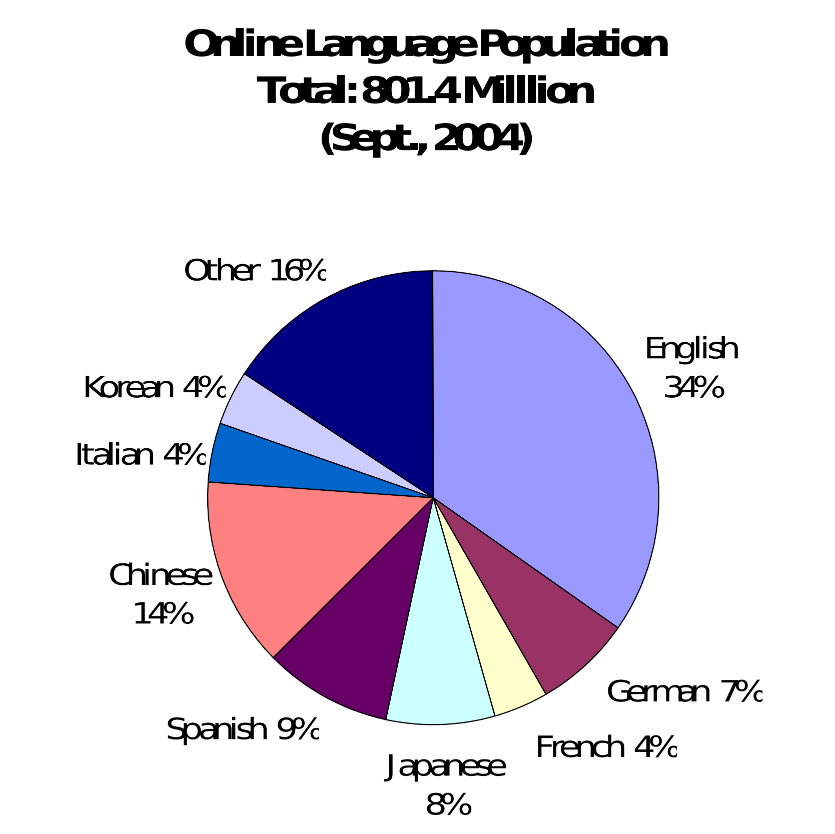
- **Grey lit** also known as **grey or hidden web** is not searchable or accessible through conventional search engines or subject directories and is not generally produced by commercial publishing organizations.
- Extension to Web

www.nla.gov.au/padi/topics/372.html

Language Barriers
Languages of the internet



http://www.netz-tipp.de/languages.html



http://www.glreach.com/globstats/index.php3

Who’s a publisher?

The Information Industry is Transforming

- Commercial businesses
- Institutional depositories
- Individual web sites
- Individuals (blogs)
- Preprint servers/networks

NASA TECHNICAL REPORTS SERVER (NTRS)

DSpace™

myspace.com a place for friends

WIKIPEDIA The Free Encyclopedia

flickr GAMMA

E-PRINT Network www.osti.gov/eprints

ELSEVIER

InfoBridge SYSTEMS

Westlaw.

CNN.com

WSJ.com THE WALL STREET JOURNAL ONLINE

CISTI - Canada Institute for Scientific and Technical Information

Google Blog Search BETA

How Dependent is Grey on the Envelope it's in?

- Literature → Founding Fathers

- Information
- Data

→ Digital Immigrants

- Digital Artifact*
- Digital Object

→ Digital Natives

*FLICC Meeting Announcement

“The definition of “technical report” may need to change. We need to recognize that some information will never be in paper format, such as modeling & simulation, interactive formulas and data streams. It was also noted that the formal structure for STI reporting is breaking down. “Technical reports” are likely to be PowerPoint presentations, preprints, journal articles or conference papers.”

- Report from the NISO Z39.18 Workshop, 30 Mar 2000

Has the internet created a new definition of grey literature: do the criteria for greyness have new meanings?

- Distributed through non-conventional channels or no commercial source of general availability
- Limited distribution – poor availability
- Non-professional or standard format
- Short life span vs. ephemeral?
- Rapid publication
- Lacks bibliographic control
- No public peer review
- Questions of authenticity & reliability
- Difficult to obtain

D

channels or no commercial source of general availability

- Internet enables self-publication and distribution
- Is it non-conventional?
- Is domain name on internet equal to publisher (ISBN/ISSN)?

Limited distribution

- Poor availability
- Technology needed to access it
- Skill to use the technology

Today’s grey literature may be a result of being too much in the public domain (grey literature of a networked world)

“But this embarrassment of riches has created a problem: How can you find anything in that mass of data?” . . . Unless you know what you’re looking for, you probably won’t find it.

- *Science* Vol. 261,13 August 1993. P.841, “Beyond Databases and E-mail”

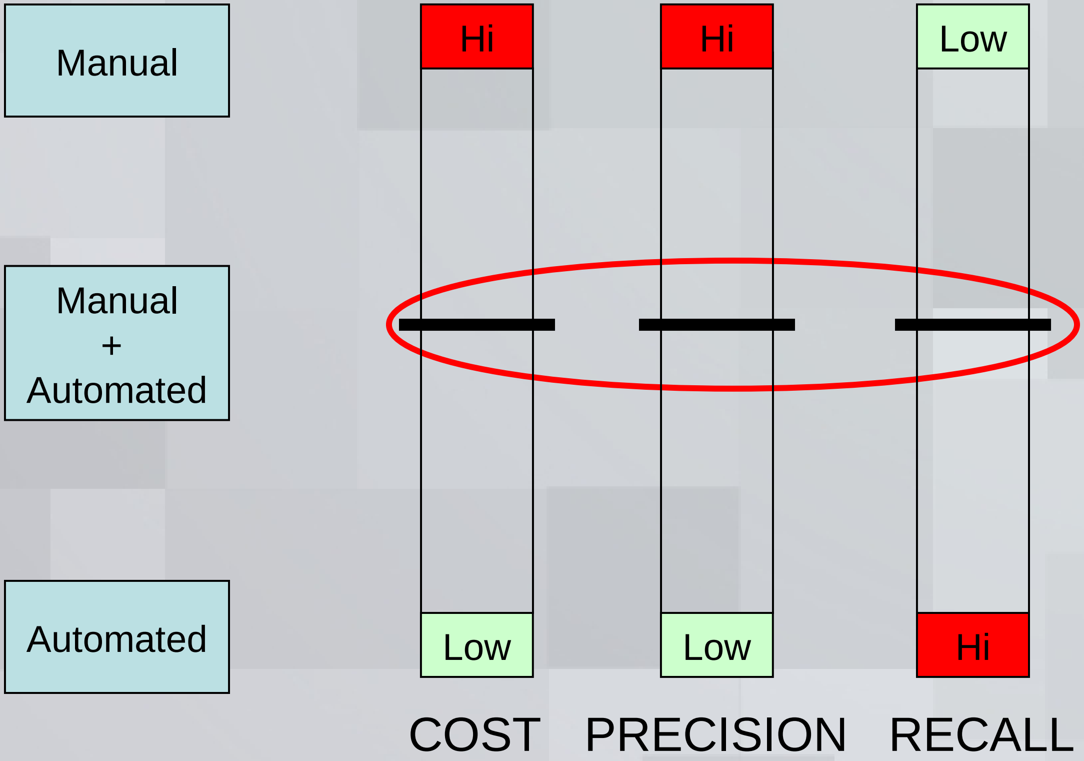
Lacks bibliographic control

- Cataloging ain’t what it used to be
- Manual vs. Machine meta data
  - – Lack of authority files
- No metadata standards
- Efforts to catalog Internet sites
  - – Cannot keep pace with the growth of this publishing medium
- New forms of organization and access

To Catalog or not to Catalog

- Always questions about GL in libraries
  - – Full cataloging → special organized collections eg. government documents → vertical files
- Today’s question is automation → manual
  - – Recall, precession, cost

Our Hypothesis



Manual

Manual
+
Automated

Automated

Hi Hi Low

Low Low Hi

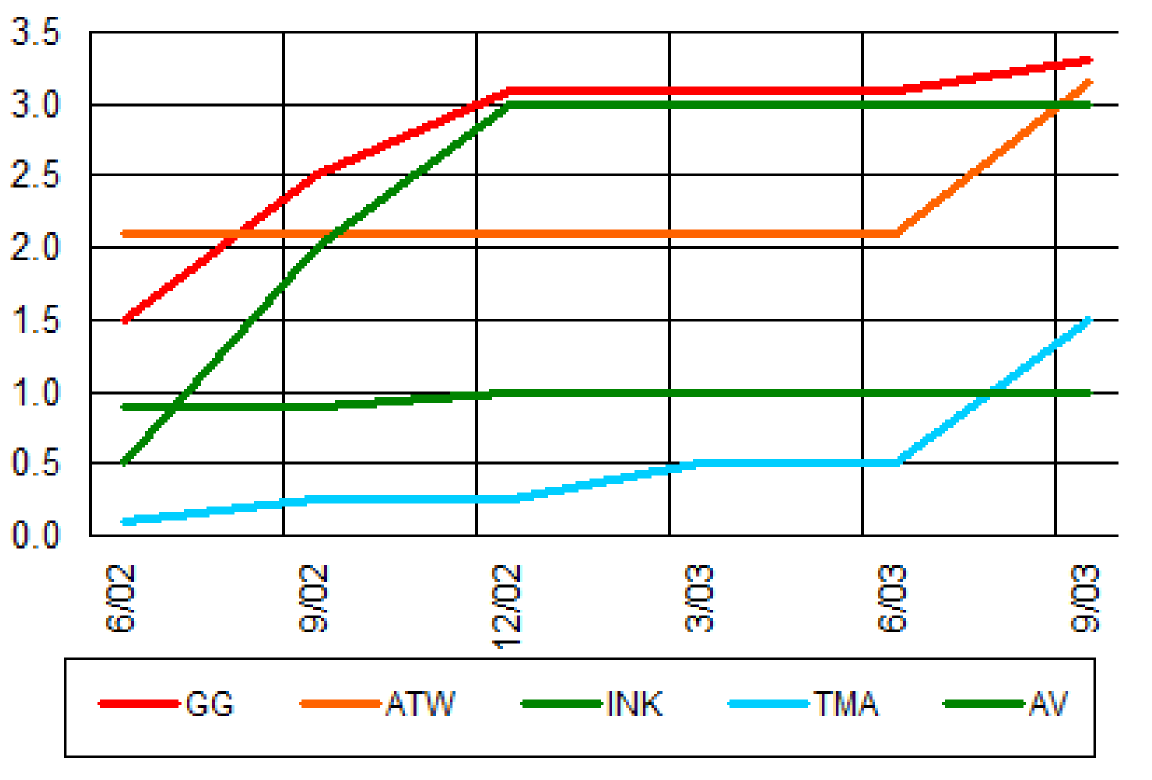
COST PRECISION RECALL

“Difficult to obtain” takes on new meaning

- Logical and physical access are interrelated
- Identification of relevant documents may be difficult and time-consuming using existing search engines
- Affected by the kinetics of Internet
  - – Documents can appear and disappear in a matter of days – documents posted to newsgroups

Focus on Search

Search Engine Size War
Search Engine Watch



http://searchenginewatch.com/showPage.html?page=sew_print&id=2156481

Search Engines & The Indexable Web

(January 2005)

11.5B pages

- Google 76%
- MSN Beta 62%
- Ask/Teoma 58%
- Yahoo! 69%

29% of indexed web (2.7Bpages) are covered by major engines

Grieli & Signorine
www.cs.uiowa.edu/~asignori/web-size/

Rapid publication

- Takes on new meaning – nearly instantaneous
- However preservation is a challenge

No public peer review

- In few instances are electronic “publications” peer-reviewed
- New quality standards for selection must be developed
- Provenance & Authenticity are key questions
- One of the most dramatic challenges epitomized by social software
- The dyke of commercial publishers
- Will the next generation see it differently?

amazon.com

YAHOO! MOVIES

Web Log (Blog) – People as Publisher

- Publicly accessible personal journal
- Frequently updated website
- Now becoming interactive
- ~2.6 M in 2003
  - 50% English
  - 50% Portuguese, Polish, Farsi, French, Spanish, German, Italian, Dutch, Icelandic

file://///Herald/www/research/projects/how-much-info-2003/internet.htm (7 of 22) [10/30/2003 4:11:59 PM]

Wiki – Community as Publisher

- Tool for collaborative authoring (publishing?)
- A type of web site that allows the visitor to easily add, remove, and some available
- Open, continuously updated
- Also can refer to the collaborative software itself

Languages of Wikipedia (11-06)

Rank	Language	Articles	Rank	Language	Articles
1.	English	1,462,910	11.	Russian	114,137
2.	German	489,585	12.	Chinese	97,981
3.	French	386,560	13.	Finnish	84,957
4.	Polish	311,145	14.	Norwegian	81,961
5.	Japanese	280,158	15.	Esperanto	60,518
6.	Dutch	237,095	16.	Slovak	57,337
7.	Italian	210,564	17.	Danish	51,776
8.	Portuguese	192,660	18.	Czech	48,820
9.	Swedish	190,931	19.	Hebrew	47,062
10.	Spanish	166,402	20.	Catalan	44,526

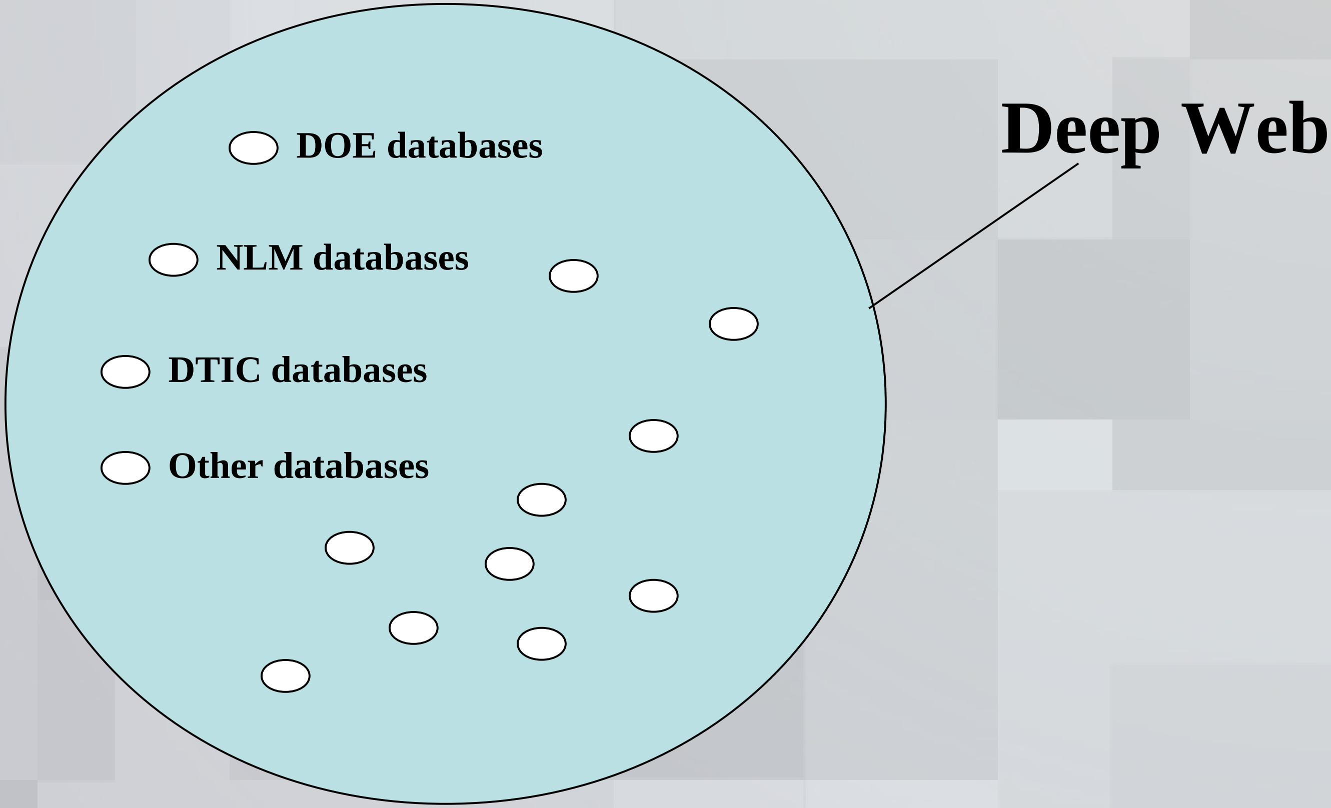
http://en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics (11/2006)

E

the Grey Barrier Challenge

- Search & Technology
  - – demand driven just in time not just in case
  - – Federated search (sci.gov) & drive to intercept ability
  - – To metadata or not to metadata
- IPR
- Quality Social Software
  - – Peer review – Public review
- Preservation & long-term access

Search & Technology
Federated Searching



Databases for CENDI agencies were scattered around the Deep Web

Slide compliments of Walt Warnick

Grey Literature & IPR

- Changes in copyright law
- Open access
- Information Commons
  - – Conventional Lit is well marked, copyrighted and clear
  - – Marking (CENDI in gov. tech reports)

Advertisement

Digital Archiving is an Urgent Challenge

- “Digital information is fragile in ways that differ from traditional technologies such as paper or microfilm. “ -- Gail Hodge, D-Lib, 2000
- More easily corruptible – Provenance becomes an issue
- Storage media shorter life spans, new generations more rapidly
- Linkages to software and hardware
- Time frame between manufacture and preservation shrinking

Questions for Discussion

Shades of Grey

- Have the lines between Grey Literature and other forms of publishing been blurred to the point that it will not be useful to distinguish?
- Do we need to redefine Grey Literature or its characteristics?
- What is the fundamental distinguishing characteristics: Hard to find or hard to get?

Vision of the Future: What's Grey?

- Star Trek Computer
- Paul Peters' "Information will be the hunter…"

In the paleo-electronic world, predator/prey relationships are rapidly reversing.

- Print world talks of “food for thought”
- Networked world users will be the prey and information is the predator

- Paul Peters
3rd NASA Foreign Acquisitions Workshop
1993