Getting statistics we can use

Judy Luther
President, Informed Strategies LLC, USA

The need for data

Both publishers and librarians are discovering that usage statistics are a topic on their agendas for several reasons:

- Librarians depend on publishers and aggregators to provide usage data
- Publisher revenues depend on budgets justified by librarians
- Librarians need this data to:
  - Obtain funding for electronic resources
  - Make internal selection decisions.

With so much information "free on the web" and declining print usage statistics reflecting the shift to electronic resources, librarians need data and new tools to "tell the story." When electronic journal usage stats are available, they confirm expectations that electronic resources delivered to the desktop are increasingly popular with users.

During the NISO Forum on Performance Measures and Statistics (http://www.niso.org/stats-rpt.html) held in Washington DC in February 2001, librarians acknowledged the need to have access to data they could use to demonstrate how the user is served and how the library supports the institution's goals. This requires both quantitative data on usage of electronic resources and qualitative data on user satisfaction.

The challenge to define the data

The White Paper on Electronic Journal Usage Statistics published by CLIR (the Council on Library and Information Resources) in October 2000, presents issues on the availability of usage data (http://www.clir.org/pubs/reports/pub94/pub94.pdf). Continued conversations emphasize the need to address the complexities of how the data is processed in order to reach meaningful conclusions about data elements.

The obvious need to reach consensus on the definition of common terms is complicated by the ways in which those activities are counted. For instance, most libraries want to count "sessions" as users without realizing that reports on sessions will vary dramatically if one publisher's system times out at 20 minutes and another publisher's system times out at 40 minutes.

Usage data comes from web logs and how this is counted also varies significantly. For instance, if one publisher counts HTML views and PDF downloads and another publisher only PDF downloads, the latter is at a distinct disadvantage. If one publisher screens out dramatic spikes from "robot contamination" and another ignores the spikes, the publisher working to present an accurate picture is at a disadvantage.

So the process of defining data that libraries would find useful is inextricably intertwined with the process of determining the protocols on how the data is collected. This must be an iterative process with librarians and experienced techies involved in the process of selecting and defining terms that represent a meaningful level of activity.

Why is this important?

When we consider how data is used, consistency and reliability are essential to being able to assess trends (comparing data over time) or assess performance (comparing data across institutions).

Data by itself is a meaningless number. It needs a context to put it in perspective. Libraries want to compare:

- this year's activity with last year's activity
- their library's performance with another library's performance
- journal title usage from one publisher with another publisher

Interpretation is contingent upon data that is produced with the same set of variables. There is growing recognition of the need to create a set of guidelines to produce data that we all can use.

Getting Statistics We Can Use

Judy Luther
Informed Strategies LLC

August 2001

Current Initiatives

- ICOLC – under revision
- NISO Forum
- ARL – eMetrics, LibQUAL
- NCLIS – definitions approved
- JISC/PALS – code of practice
- Elsevier White Paper
- STM – validation and compliance

Council on Libraries & Information Resources

- Electronic Journal Usage Statistics

  www.clir.org/pubs/reports/pub94/pub94.pdf

  - Libraries: Ohiolink, Los Alamos, FCLA
  - Publishers: Academic, Elsevier, MCB, IOP
  - Aggregators: Jstor, Catchword, Highwire

Informed Strategies

Challenges - emotional

Fear of statistics

- **Unclear what to count**
  - **User activities**
  - **Institutional summary**
  - **Journals/database**
- **Unclear what it means**
  - **May use for cancellation**
  - **Use does not equal value**
  - **Denominator is the key***

Informed Strategies

Challenges - functional

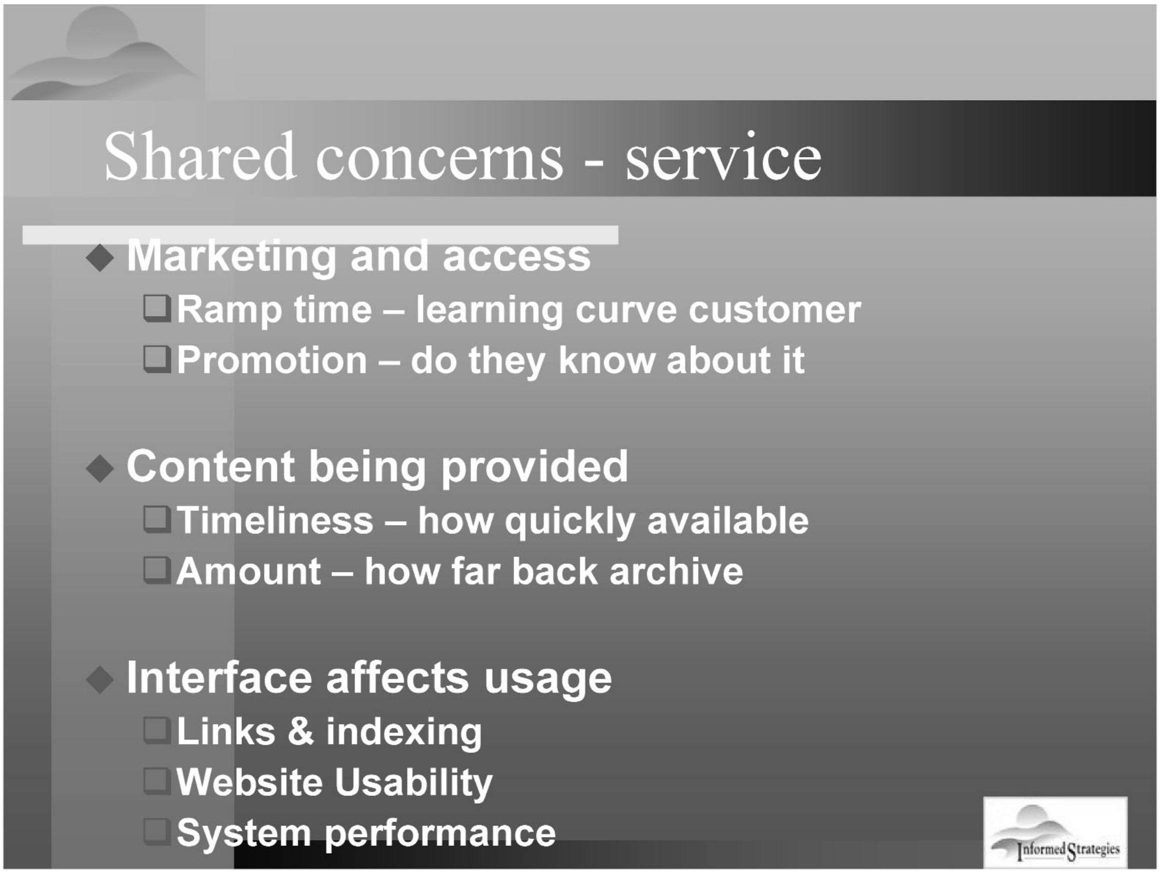
- Reliability of data*
  - Caching – how to include
  - Robot contamination
  - Software – how it counts
  - Log files – double clicks, images
- Process for development
  - Learning curve - evolutionary
  - Libraries must depend on publishers
  - What is needed? Possible? Useful?

Informed Strategies

Shared concerns - data

- Lack of comparable data
  - Varied definitions and implementation
  - Only valid from same platform*
- Lack of context
  - No historical reference
  - Activity does not equate to use
- Incomplete usage data
  - Multiple sources and formats

Informed Strategies


Shared concerns - service
Marketing and access
Ramp time – learning curve customer
Promotion – do they know about it
Content being provided
Timeliness – how quickly available
Amount – how far back archive
Interface affects usage
Links & indexing
Website Usability
System performance
Informed Strategies


Shared concerns - issues
Interpretation
Cost per use
How to assess value
User privacy
Data on users & on libraries
FTC in US, European Commission
Informed Strategies

Publisher Issues

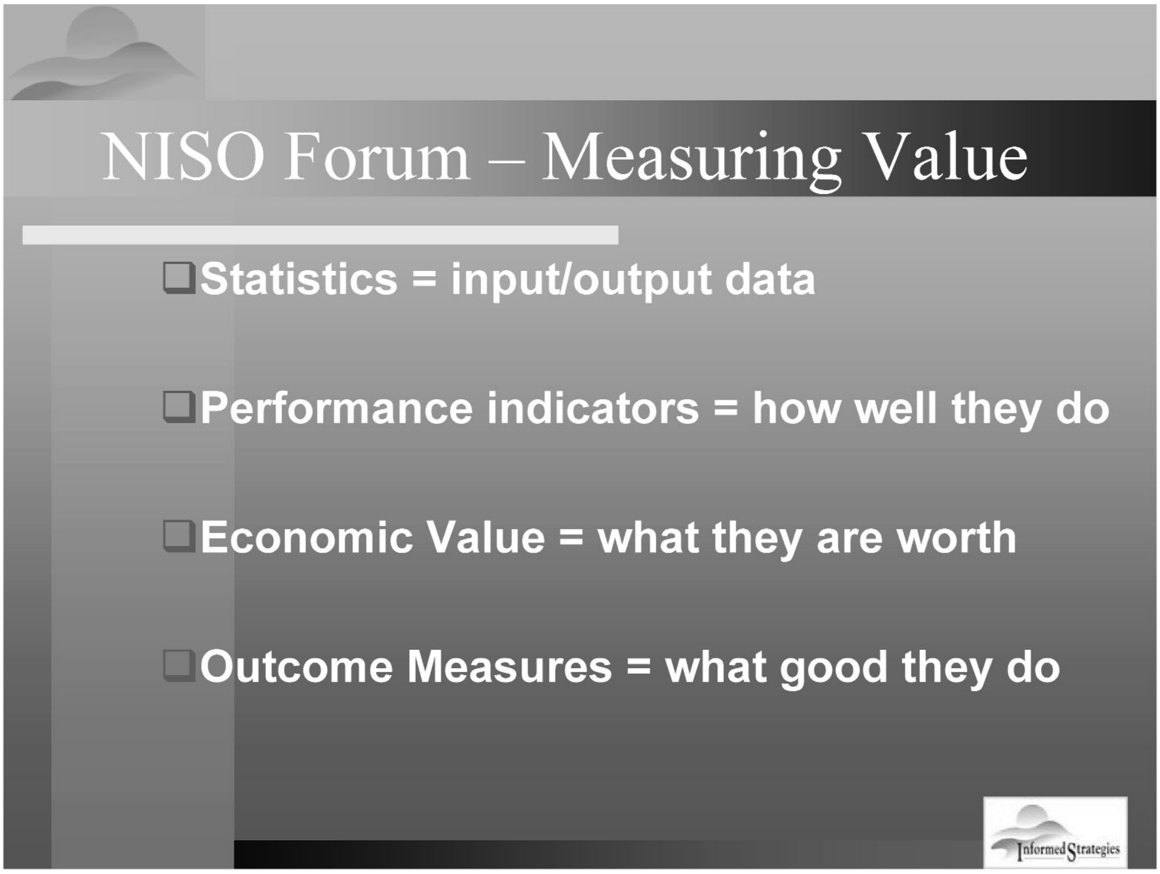
- Internal applications
- Expensive to create functionality
- New role for publishers

Informed Strategies

Library Issues

- "To Know" - manage internally
  - Make acquisition/space/service decisions
  - Rethink ownership vs access
  - Journals = 70% budget
- "To Show" – manage externally
  - Budget justification
  - Demand for journals not owned (Ohiolink)
  - Survival

Informed Strategies


NISO Forum – Measuring Value
Statistics = input/output data
Performance indicators = how well they do
Economic Value = what they are worth
Outcome Measures = what good they do
Informed Strategies


Measures Vary by Product Type
Bibliographic data – searches
Journals – downloads
Books & reference tools – data bits
Informed Strategies

What are we measuring?

- What is being used?
  - Data by journal title
- Who is using it?
  - Use of IP addresses
- How is the database used?
  - Searching by subject vs browsing from TOC
- When is the content being used?
  - Administrative

Informed Strategies

Producing statistics

1. Identifying data – defining elements
2. Analyzing data – assigning meaning
3. Presenting/delivering – make available

Informed Strategies

How will the data be presented?

- Query the database, specify period
- Two years of data up online
- Downloadable
- Compare with selected institutions
- Email alert

Informed Strategies

PALS Working Group

1. Gateways/hosts
2. Defining searches/sessions
3. Authentication – proxy servers
4. Market elements – institution
5. Data Model – taxonomy
6. Data Processing – data integrity
7. Types of reports

Informed Strategies


Opportunity for the Community
Not a point of differentiation*
Valid reliable data
Working together - iterative*
Best practice
Global issue
Informed Strategies