MESUR: overview and lessons learned

Johan Bollen
Digital Library Research & Prototyping Team
Los Alamos National Laboratory - Research Library

jbollen@lanl.gov

Acknowledgements:
Herbert Van de Sompel (LANL), Marko A. Rodriguez (LANL), Ryan Chute (LANL), Lyudmila L. Balakireva (LANL), Luis Bettencourt (LANL), Aric Hagberg (LANL)

Research supported by the Andrew W. Mellon Foundation.

Usage data has arrived.

Value of usage data/statistics is undeniable:

- Business intelligence
- Scholarly assessment
- Monitoring of scholarly trends
- Enhanced end-user services

Production chain:

- Recording
- Aggregation
- Analysis

Challenges and opportunities at all links in chain.

1. Recording: requirements, data models, standardization
2. Aggregation: standards, sampling, ownership
3. Analysis: frequentist (usage statistics) vs. structural (usage data), dependent on (1) and (2)

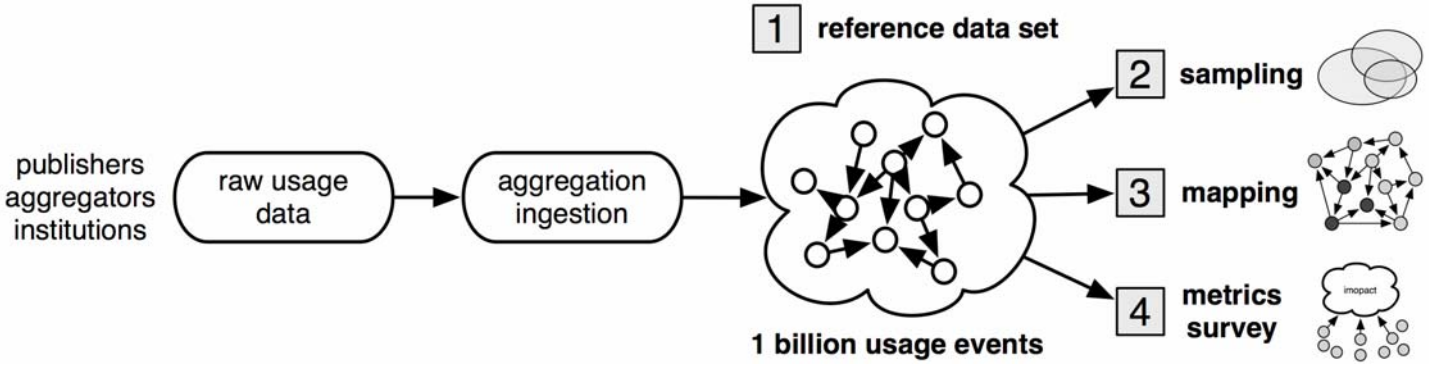
The MESUR project.

“The Andrew W. Mellon Foundation has awarded a grant to Los Alamos National Laboratory (LANL) in support of a **two-year project** that will **investigate metrics** derived from the network-based **usage** of scholarly information. The Digital Library Research & Prototyping Team of the LANL Research Library will carry out the project.
The project's major objective is **enriching** the toolkit used for the assessment of the impact of scholarly communication items, and hence of scholars, with **metrics that derive from usage data**.”

MESUR: A scientific approach to scholarly assessment

Investigate validity of usage data and usage-based metrics

1. **Exploring many possibilities, many facets of impact: Not selling 1 metric!**
2. Cross-validation: compare to existing, accepted **journal**-focused metrics and data
3. Explorative approach: not top-down, bottom-up exploration
4. Focus on
    1. Usage **data**, not usage statistics
    2. **Network** metrics, not descriptive statistics



How to obtain 1,000,000,000 usage events?

Politely asked selected publishers, aggregators and institutional consortia for usage data.

Agreements achieved throughout 2006-2007. Stringent privacy requirements.

1. Publishers: Six major international publishers
2. Aggregators: Four significant international aggregators
3. Institutions: Four large USA-based university consortia

Present scale of the MESUR DB:

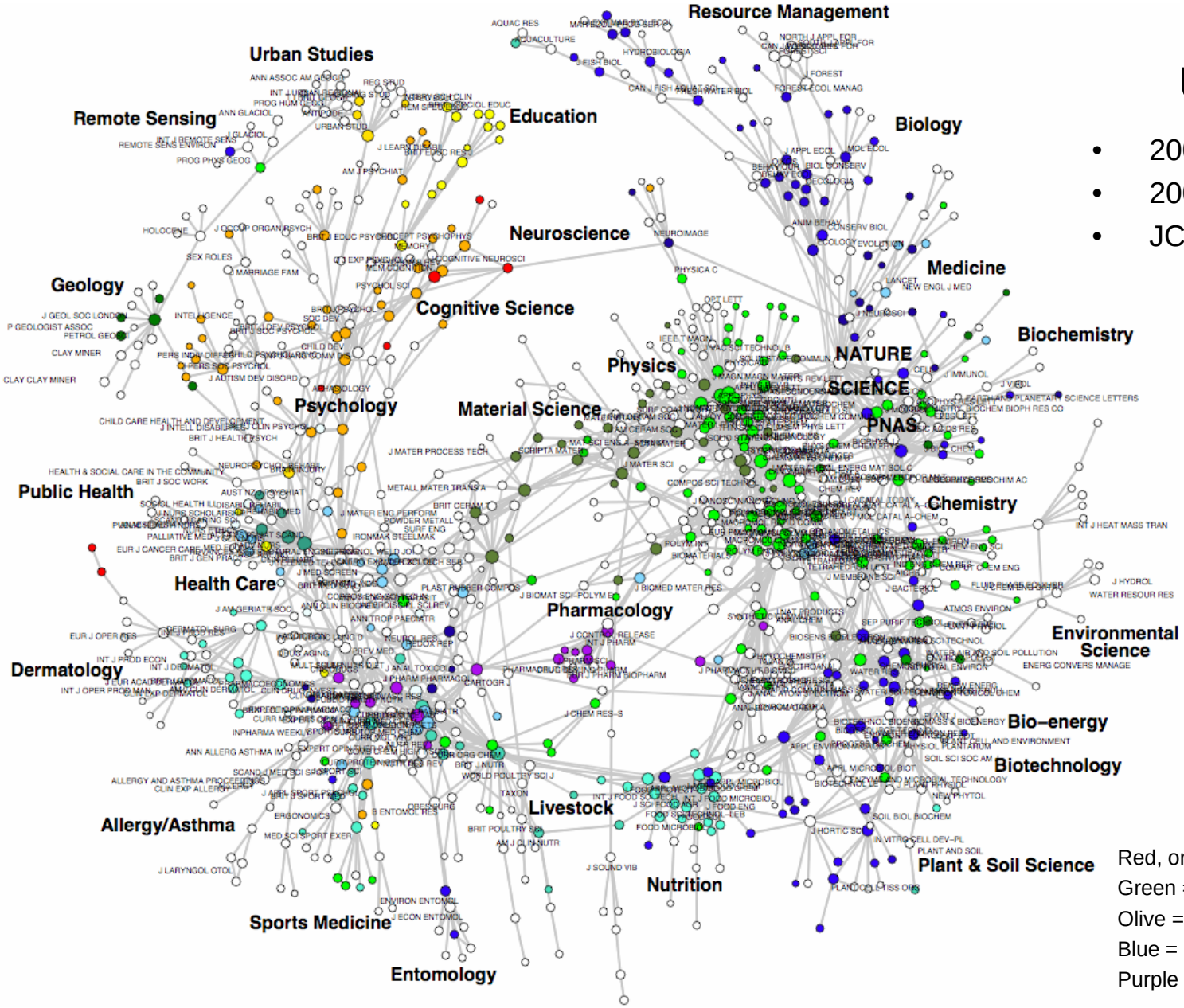
- +-1,000,000,000 usage events and +500,000,000 citations
- 2002-2007, but mostly 2006
- > 50M documents
- +- 100,000 serials (incl. newspapers, magazines,…)
- COUNTER reports from 2 publishers for nearly 2,000 institutions

MESUR's usage data

Requirements:

- Article level usage events
- Fields: session ID, date/time, document ID and/or metadata, request type
- Note difference with COUNTER-like usage statistics: sequence or clickstream is preserved!

```
2007, 9, 1, 0, 0, 1 CFA    cffoe   A172080.N1.Vanderbilt.Edu       unknown AST    A      1996SPIE.2828..64S    http://foe.edu/abs/199
2007   9     1     0     0     1     CFA    cffoe   210.94.41.89    unknown PHY    A      2007ApPhL.90a2120C    http://foe.edu/abs/2
2007   9     1     0     0     1     CFA    cffoe   24-196-228-125.dhcp.gwnt.ga.charter.com unknown AST    A      2000ASPC.213.3
2007   9     1     0     0     4     CFA    cffoe   163.152.35.114  4700387eae     PHY    A      1993WRR..29.133S      http://foe.edu/
20079     1     0     0     6     CFA    cffoe   pd9e980fc.dip0.t-ipconnect.de  45f0c69881     AST    X      2007AN..328.841H      htt
2007   9     1     0     0     1     CFA    cffoe   A172080.N1.Vanderbilt.Edu       unknown AST    A      1996SPIE.2828..64S    http
```



Usage map

- 200M usage events
- 2006 usage only
- JCR journals (+-7600)

Red, orange= psych, cogn
Green = phys, chem
Olive = material science
Blue = biology
Purple = pharma

Digital Library Research & Prototyping Team
Research Library, Los Alamos National Laboratory
@ JISC workshop, Humboldt U. zu Berlin, 2008

Examples of **usage network** rankings

2004 Impact Factor

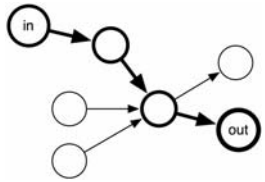
	value	journal
1	49.794	CANCER
2	47.400	ANNU REV IMMUNOL
3	44.016	NEW ENGL J MED
4	33.456	ANNU REV BIOCHEM
5	31.694	NAT REV CANCER

Pagerank

	value	journal
1	0.0016	SCIENCE
2	0.0015	NATURE
3	0.0013	PNAS
4	0.0010	LNCS
5	0.0008	J BIOL CHEM

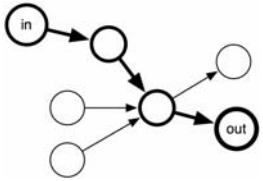
betweenness

	value	journal
1	0.035	SCIENCE
2	0.032	NATURE
3	0.020	PNAS
4	0.017	LNCS
5	0.006	LANCET



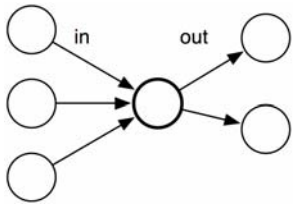
Closeness

	value	journal
1	0.670	SCIENCE
2	0.665	NATURE
3	0.644	PNAS
4	0.591	LNCS
5	0.587	BIOCHEM BIOPH RES CO



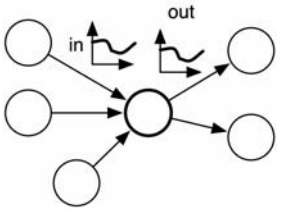
In-Degree

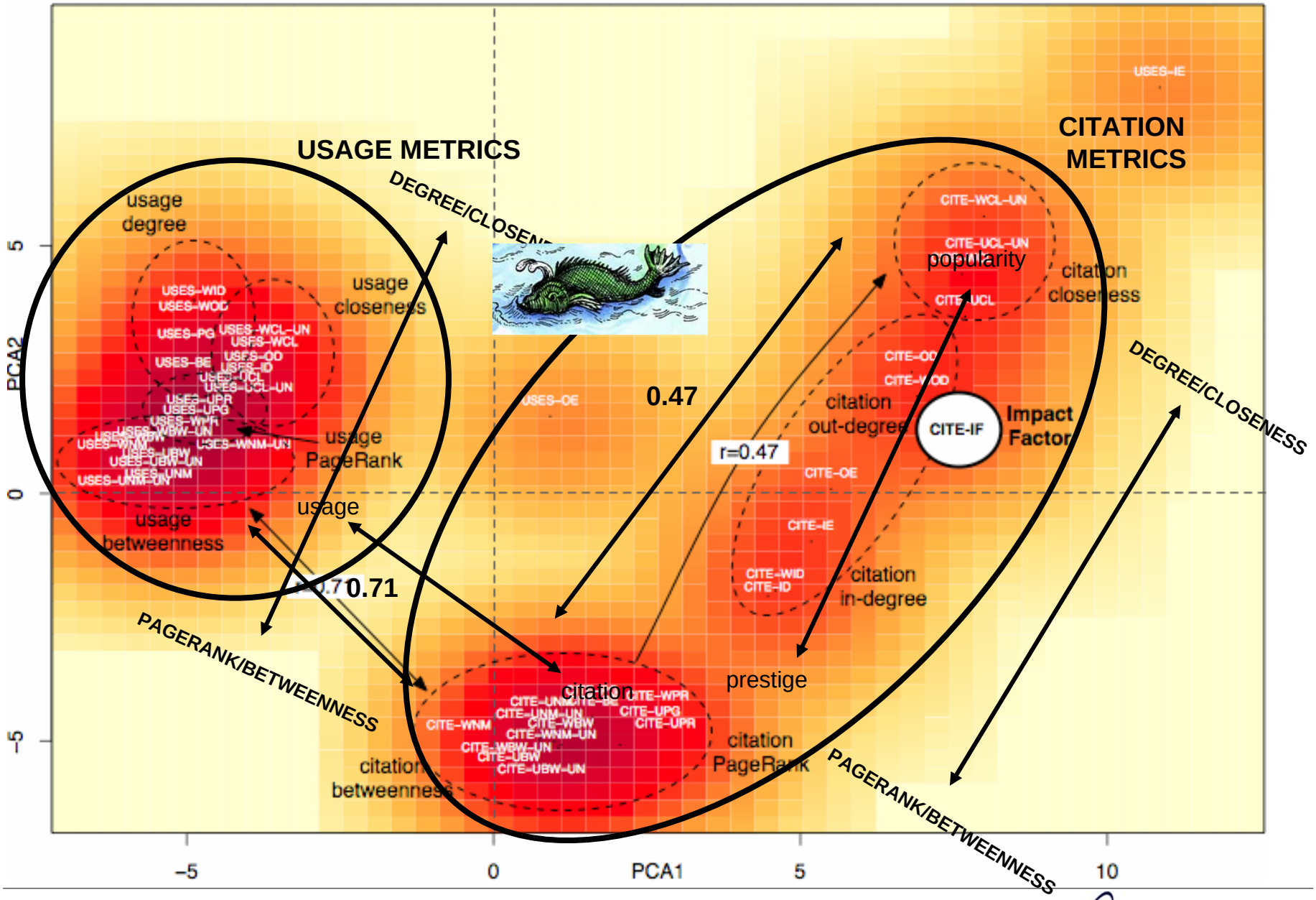
	value	journal
1	4195	SCIENCE
2	4019	NATURE
3	3562	PNAS
4	2438	J BIOL CHEM
5	2432	LNCS



In-degree entropy

	Value	journal
1	9.364	MED HYPOTHESES
2	9.152	PNAS
3	9.027	LIFE SCI
4	8.939	LANCET
5	8.858	INT J BIOCHEM CELL B




USAGE METRICS
CITATION METRICS
usage degree
usage closeness
usage PageRank
usage betweenness
usage
DEGREE/CLOSENESS
PAGERANK/BETWEENNESS
0.47
0.71
r=0.47
popularity
citation closeness
citation out-degree
CITE-IF
Impact Factor
citation in-degree
prestige
citation PageRank
citation betweenness
citation
DEGREE/CLOSENESS
PAGERANK/BETWEENNESS
PCA1
PCA2
-5
0
5
10

COUNTER reports: information loss

From: **www.niso.org/presentations/MEC06-03-Shepherd.pdf**

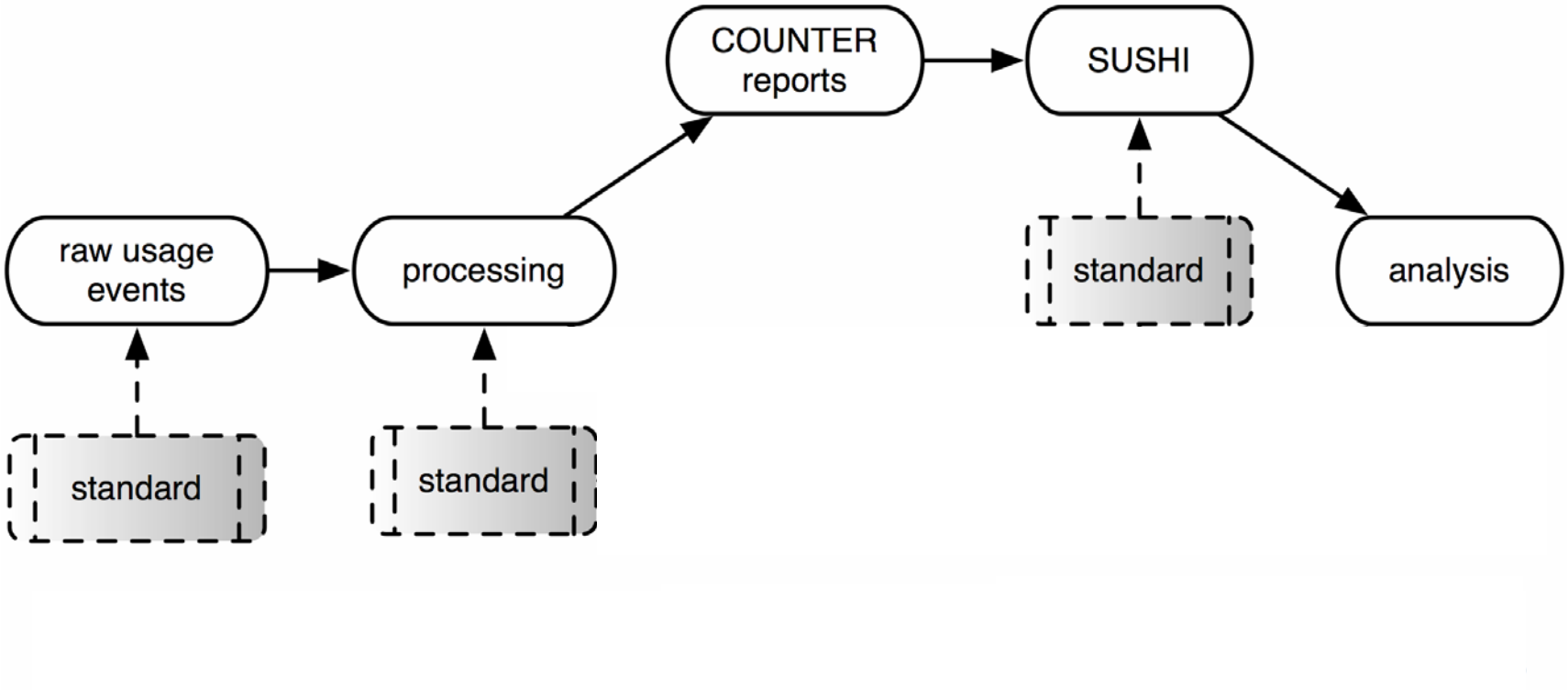
QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

50 years of network science:

- Mapping
- Social network metrics

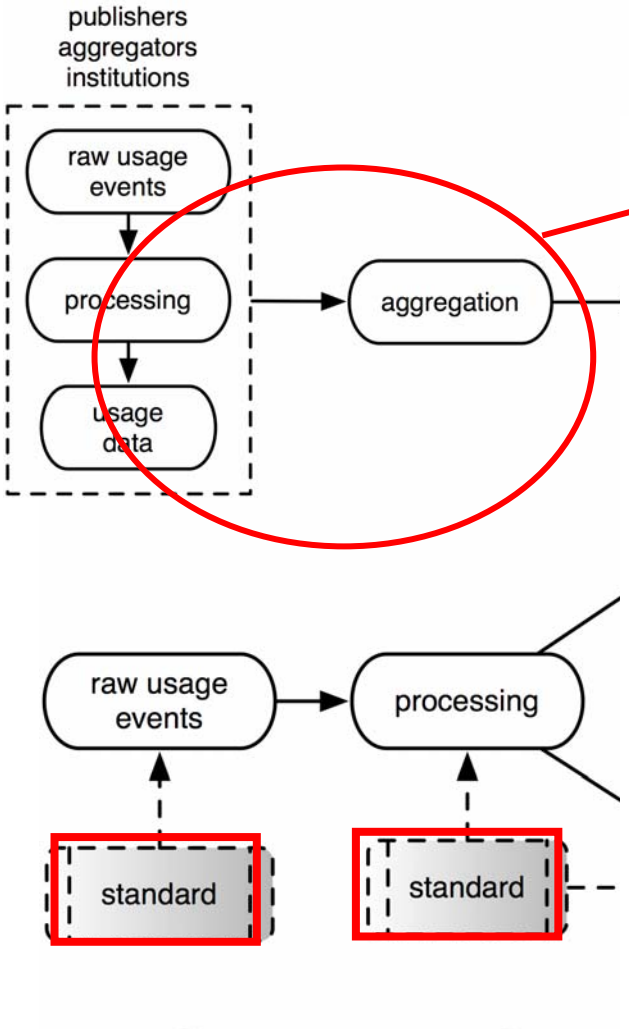
1. **Event ID:** distinguish usage events ?
2. **Referent ID:** DOI, SICI, metadata
3. **User/Session ID:** define groups of events related by user?
4. **Date and time ID:** identify data and time of event
5. **Request types**: identiby type of request issued by user

We have COUNTER/SUSHI. How about the aggregation of item-level usage data?



If there is value in aggregating COUNTER and other reports, there is considerable value in aggregating item-level usage data.

We need a standardized representation framework for item-level usage data.



MESUR: Ad hoc parsing is highly problematic

- field semantics
- field relations
- various data models

Framework objectives:

1. Minimize data loss:
    1. Preserve event info
    2. Preserve sequence info
    3. Preserve document metadata
2. **"realistic"**: scalability and granularity.
3. Apply to variety of usage data, i.e. **no inherent bias** towards specific type of usage data

Requirements for usage data representation framework.

Needs to minimally represent following concepts:

1. **Event ID:** distinguish usage events
2. **Referent ID:** DOI, SICI, metadata
3. **User/Session ID:** define groups of events related by user
4. **Date and time ID:** identify data and time of event
5. **Request types**: identiby type of request issued by user

Implications

1. **Sequence**: session ID and date/time preserves sequence
2. **Privacy:** session ID groups events not by user ID
3. **Request types:** filter on types of usage

Aggregating item-level usage data requires standardized aggregation framework.

Standardization objectives similar to work done for COUNTER and SUSHI:

1. **Serialization (~COUNTER):**
   - standard to serialize usage data
   - Suitable for large-scale, open aggregation
   - Event provenance and identification
2. **Transfer protocol (~SUSHI):**
   - Communication of usage data between log archive and aggregator
   - Allow open aggregation across stakeholders in scholarly community
3. **Privacy standard:**
   - Standards need to address privacy concerns
   - Should allow emergence of trusted intermediaries: aggregation ecology

LANL has made proposals based on community standards

OpenURL ContextObject to represent usage data

Event information:
*** event datetime**
*** globally unique event ID**

Referent
*** identifier**
*** metadata**

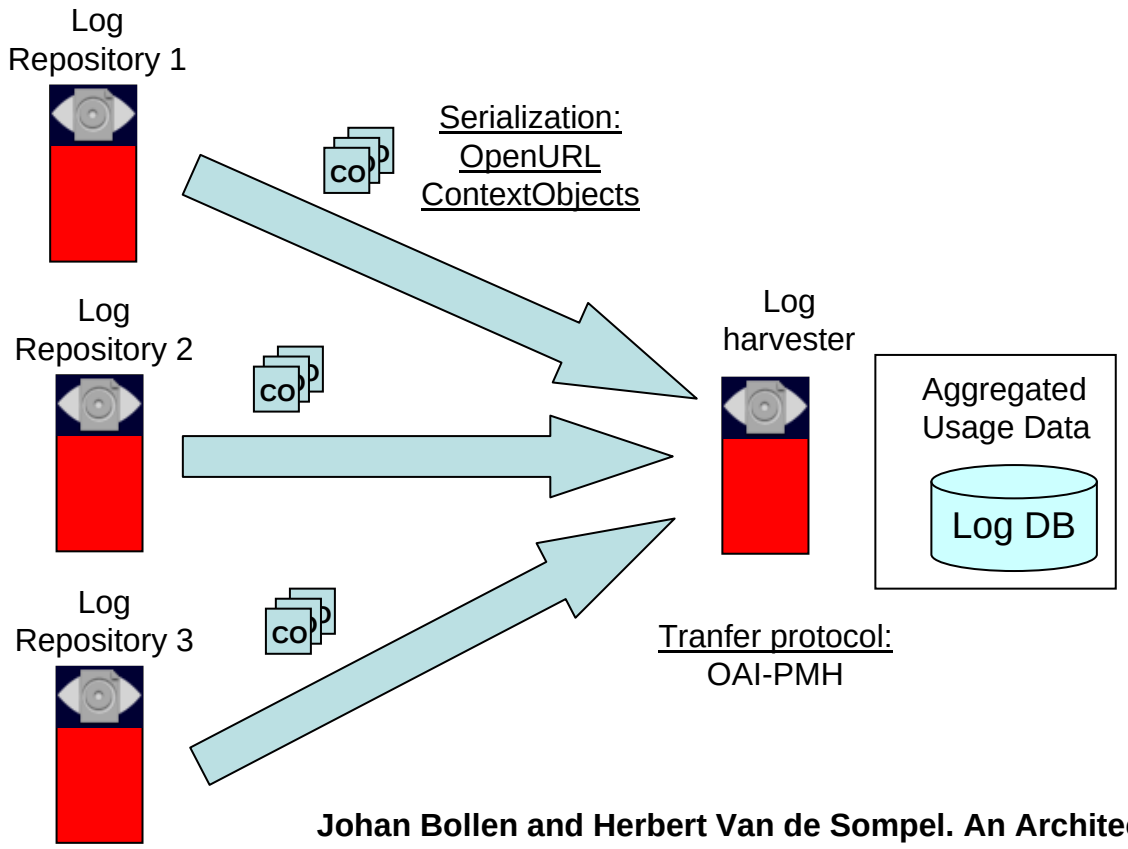
Requester
*** User or user proxy: IP, session, …**

ServiceType

Resolver:
*** identifier of linking server**

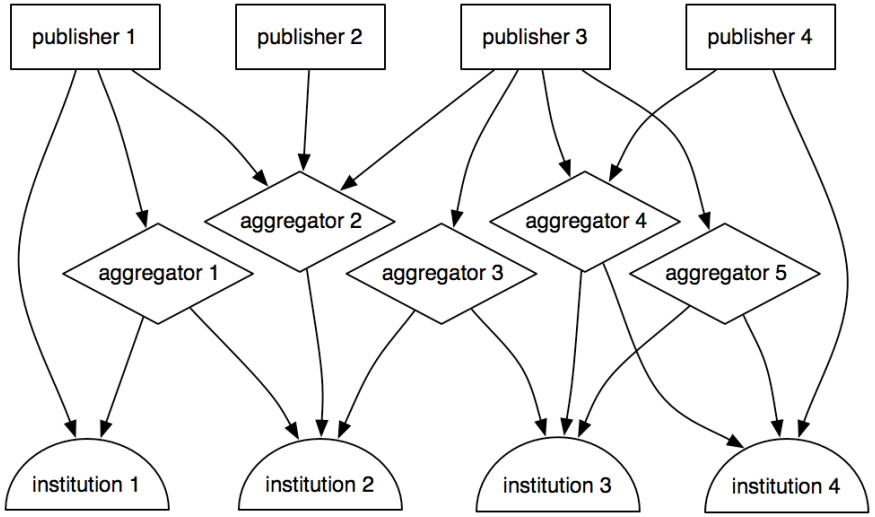
```
<?xml version="1.0" encoding="UTF-8"?>
<ctx:context-object
 timestamp="2005-06-01T10:22:33Z" …
 identifier="urn:UUID:58f202ac-22cf-11d1-b12d-002035b29062" …>
…
<ctx:referent>
 <ctx:identifier>info:pmid/12572533</ctx:identifier>
 <ctx:metadata-by-val>
  <ctx:format>info:ofi/fmt:xml:xsd:journal</ctx:format>
  <ctx:metadata>
   <jou:journal xmlns:jou="info:ofi/fmt:xml:xsd:journal"> …
   <jou:atitle>Isolation of common receptor for coxsackie B …
   <jou:jtitle>Science</jou:jtitle>
…
</ctx:referent>
…
 <ctx:requester>
   <ctx:identifier>urn:ip:63.236.2.100</ctx:identifier>
 </ctx:requester>
…
 <ctx:service-type>
  …
  <full-text>yes</full-text>
  …
 </ctx:service-type>
 …
  Resolver…
  Referrer…
  ….
</ctx:context-object>
```

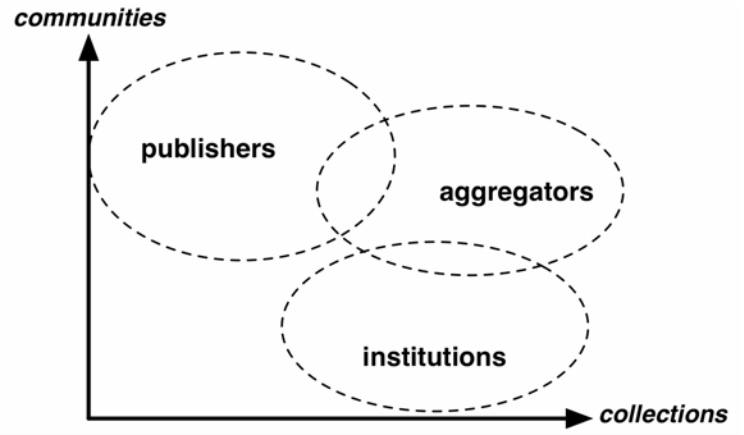
Aggregation framework: existing standards



Johan Bollen and Herbert Van de Sompel. An Architecture for the aggregation and analysis of scholarly usage data. In Joint Conference on Digital Libraries (JCDL2006), pages 298-307, June 2006.

Related to sampling, ownership and copyright:





Main players:

- **Individual institutions**
- **Aggregators**
- **Publishers**

Each represent different, and possibly overlapping, samples of the scholarly community.

Institutions:

- Institutional communities
- Many collections

Aggregators:

- Many communities
- Many collections

Publishers:

- Many communities
- Publisher collection

Some relevant publications.

Johan Bollen, Herbert Van de Sompel, and Marko A. Rodriguez. **Towards usage-based impact metrics: first results from the MESUR project.** In Proceedings of the Joint Conference on Digital Libraries, Pittsburgh, June 2008

Marko A. Rodriguez, Johan Bollen and Herbert Van de Sompel. **A Practical Ontology for the Large-Scale Modeling of Scholarly Artifacts and their Usage,** In Proceedings of the Joint Conference on Digital Libraries, Vancouver, June 2007

Johan Bollen and Herbert Van de Sompel. **Usage Impact Factor: the effects of sample characteristics on usage-based impact metrics.** (cs.DL/0610154)

Johan Bollen and Herbert Van de Sompel. **An architecture for the aggregation and analysis of scholarly usage data.** In Joint Conference on Digital Libraries (JCDL2006), pages 298-307, June 2006.

Johan Bollen and Herbert Van de Sompel. **Mapping the structure of science through usage.** Scientometrics, 69(2), 2006.

Johan Bollen, Marko A. Rodriguez, and Herbert Van de Sompel. **Journal status.** Scientometrics, 69(3), December 2006 (arxiv.org:cs.DL/0601030)

Johan Bollen, Herbert Van de Sompel, Joan Smith, and Rick Luce. **Toward alternative metrics of journal impact: a comparison of download and citation data.** Information Processing and Management, 41(6):1419-1440, 2005.