The MESUR Project
In Search of Usage-Based Metrics of Scholarly Impact

http://www.mesur.org

Johan Bollen [1] and Herbert van de Sompel [2]
Digital Library Research & Prototyping Team
Los Alamos National Laboratory - Research Library

[1] jbollen@lanl.gov
http://public.lanl.gov/jbollen
[2] herbertv@lanl.gov
http://public.lanl.gov/herbertv

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

Research supported by the Andrew W. Mellon Foundation.

Scholarly communication in the digital age

Herbert's good old provocative statement:

The current digital scholarly communication system is a mere scanned copy of its paper-based ancestor.

- Interpret: A lot still needs to change for us to achieve a genuine digital scholarly communication system.

- The Digital Library Research & Prototyping Team at the Los Alamos National Laboratory researches various aspects of scholarly communication in the digital age:
  - Interoperability
  - Digital Preservation
  - Repository Architecture
  - Peer-review
  - …
  - Metrics

MESUR is Paradigm Shift Material

MESUR looks into one aspect that would benefit from some change: assessment of scholarly impact

- The Thomson Scientific IF was about the only metrics that could be computed in a paper-based era.
- But we don't live in the paper-based era anymore. So MESUR researches metrics for the digital era:
  - Usage-based metrics:
    - Access to scholarly materials happens via networked systems, not via paper stored in libraries.
    - Networked systems can record a great deal about access to materials; much more than libraries could about access to paper.
  - Network-based metrics:
    - Scholarly communication generates networks, e.g. citation networks, co-authorship networks, usage networks, …
    - A wide variety of metrics can be computed for such networks; much more than simple citation counts.

The Promise of Usage Data

Metrics based on usage data have significant potential in the digital era

- Can be recorded for all digital scholarly content, i.e. papers, journals, preprints, blog postings, datasets, chemical structures, software, …
  - Not just for ~ 10,000 journals
- Is recorded starting immediately after *publication*
  - Not once read and cited (think publication delays)
  - Rapid indicator of scholarly trends

- So the interest in usage data from projects such as COUNTER, Citebase, IKS and MESUR should not come as a surprise!

The Promise of Usage Data

LANL	Usage PR	IF (2003)	Title (abbv.)
1	60.196	7.035	PHYS REV LETT
2	37.568	2.950	J CHEM PHYS
3	34.618	1.179	J NUCL MATER
4	31.132	2.202	PHYS REV E
5	30.441	2.171	J APPL PHYS

CSU	Usage PR	IF (2003)	Title (abbv.)
1	78.565	21.455	JAMA-J AM MED ASSOC
2	71.414	29.781	SCIENCE
3	60.373	30.979	NATURE
4	40.828	3.779	J AM ACAD CHILD PSY
5	39.708	7.157	AM J PSYCHIAT

MSR	Usage PR	IF (2005)	Title (abbv.)
1	15.830	30.927	SCIENCE
2	15.167	29.273	NATURE
3	12.798	10.231	PNAS
4	10.131	0.402	LECT NOTES COMP SCI
5	8.409	5.854	J BIOL CHEM

- The LANL usage-based metric clearly reflects which journals are important to LANL.
- The CSU usage-based metric reflects general importance of certain journals and some local preferences.
- If enough usage data is aggregated, will the derived metrics be globally representative?

And the Obvious Challenges of Usage Data

Usage data comes with significant challenges

- What exactly is usage?
  - E.g. various types of usage (download pdf, email abstract, …); impact of user interface on usage recordings, …
- Privacy concerns
- Aggregating item-level usage data across networked systems:
  - Standardized recording
  - Standardized aggregating
  - Click-streams across networked systems
- How to deal with bots?

Network-Based Metrics

We have 50 years of network science available to us

- A wide variety of metrics has been proposed to characterize networks, and to assess the importance of nodes in a network
  - E.g. social network analysis, small world graphs, graph theory, social modeling
- So when defining metrics for scholarly communication (clearly a network), we should probably leverage network science
  - Cf. Google's PageRank versus Alta Vista's statistical ranking
  - Cf. Most selling CD (Britney Spears) versus most influential CD
- A network (and hence a network-based metric) takes context into account; a statistical count does not.
- Readings:
  - Barabasi (2003) Linked.
  - Wasserman (1994). Social network analysis.

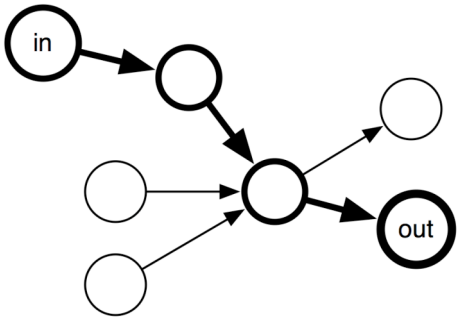
Network-Based Metrics

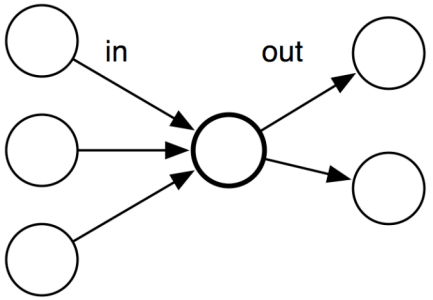
Classes of metrics:

- Degree
- Shortest path
- Random walk
- Distribution

Shortest path

- Closeness
- Betweenness
- Newman



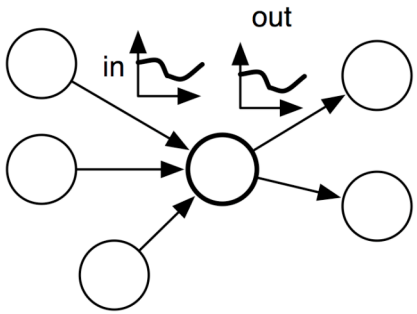


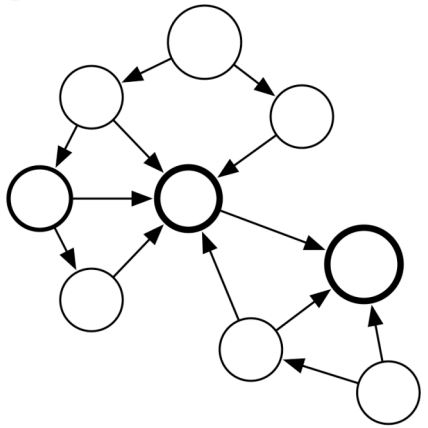
Degree

- In-degree
- Out-degree

Distribution

- In-degree entropy
- Out-degree entropy
- Bucket Entropy





Random walk

- PageRank
- Eigenvector

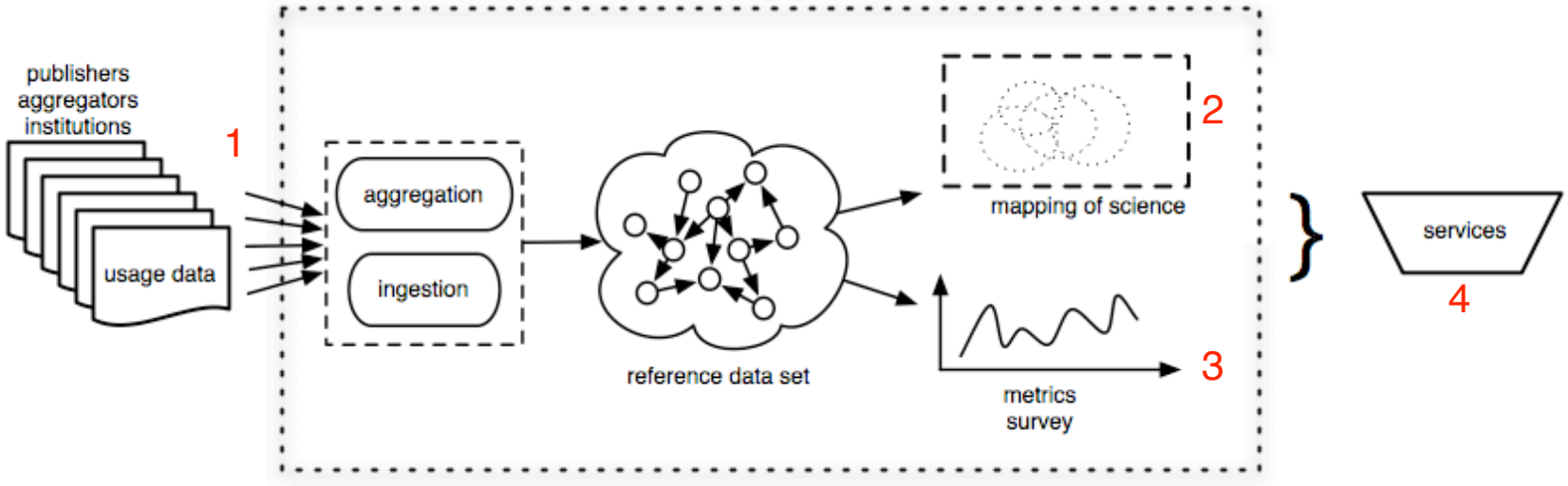
MESUR: A Thorough, Scientific Approach

We are not kiddin'

1. Create very large-scale reference data set
   a) Usage, citation and bibliographic data combined
   b) Various communities, various collections
2. Investigate sampling issues:
   a) Effects of sampling on usage-based assessment
   b) Uncertainty quantification: noise, bots, …
3. Investigate validity of usage data and usage-based metrics
   a) Cross-validation: compare to other journal metrics, e.g. citation-based IF
   b) Not selling 1 metric: exploring many possibilities, many facets of impact

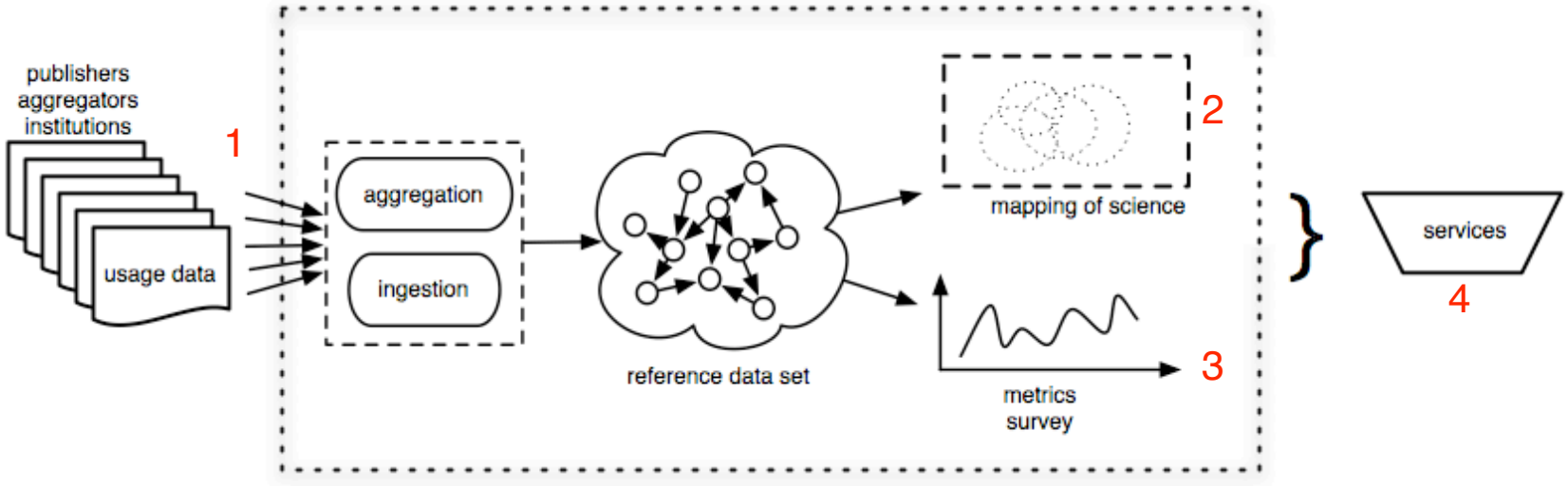
MESUR: Project Phases

1) Usage data acquisition
2) Science mapping from usage graphs
3) Metrics survey
4) Services



MESUR: Project Phases

1) Usage data acquisition
2) Science mapping from usage graphs
3) Metrics survey
4) Services



How to Obtain 1,000,000,000 Usage Events?

Politely ask publishers, aggregators, institutions

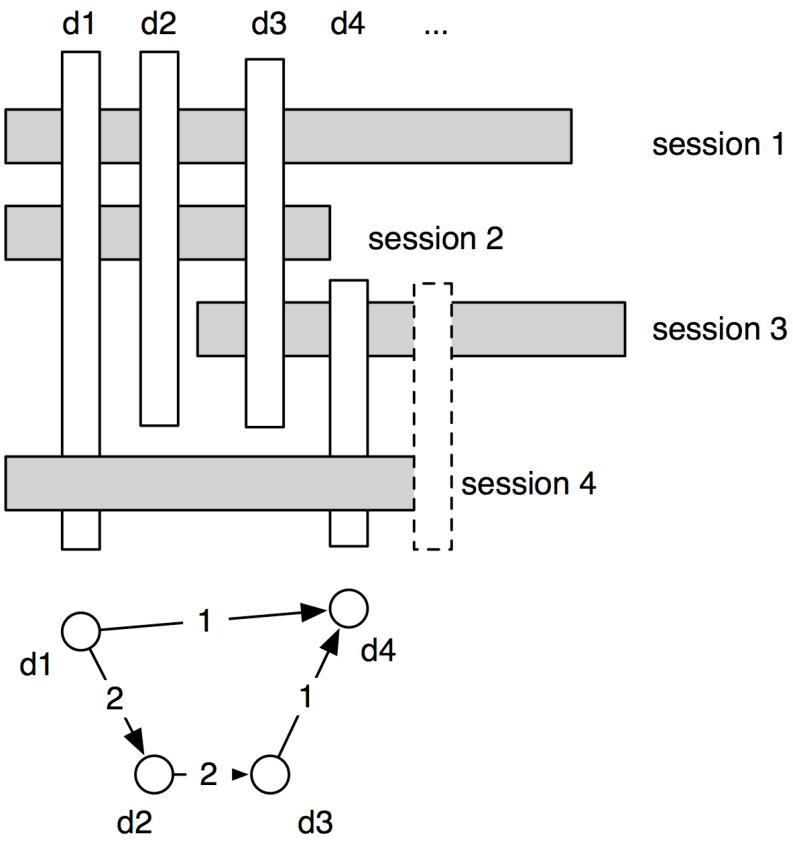
- Scale: > 1,000,000,000 usage events
- Period: 2002-2007, but mostly 2006
- Span:
  - > 50M documents
  - > 100,000 journals (inc. newspapers, magazines,…)

Some Minimal Requirements for Usage Data

In order to be able to construct usage-based networks

- Article level usage events
- Fields: unique session ID, date/time, unique document ID and/or metadata, request type

Generating a Network from Usage Data



Same session ~ documents relatedness

- Same session, same user: common interest
- Frequency of co-occurrence = degree of relationship
- Normalized: conditional probability

Usage data is on article level:

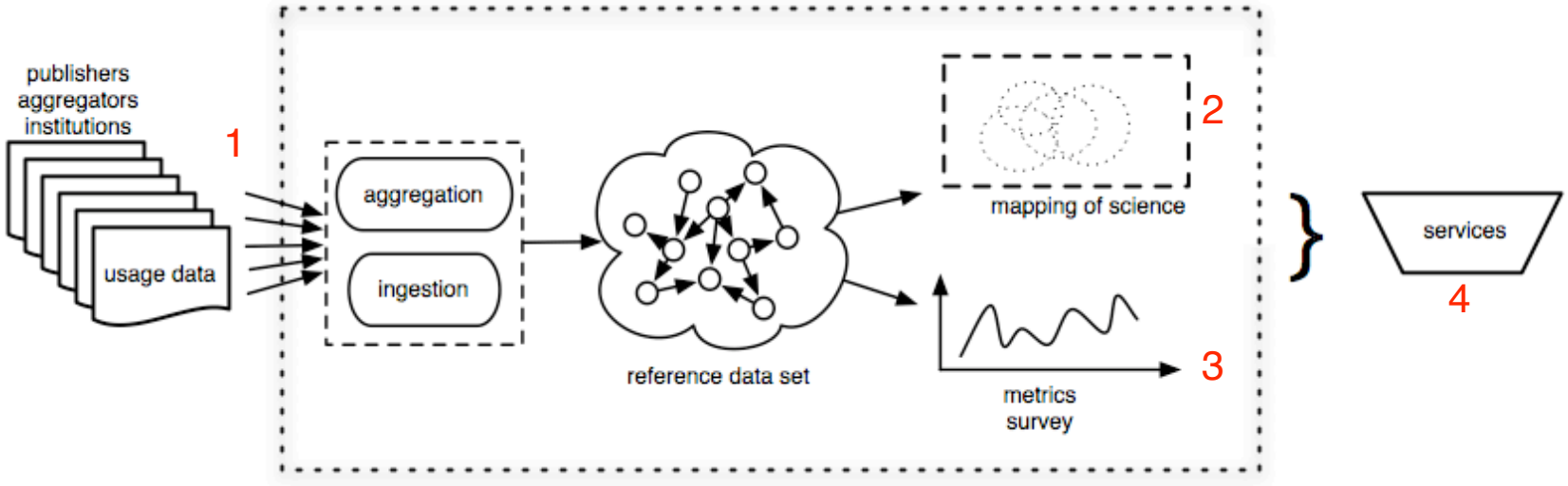
- Works for journals **and** articles
- **Anything** for which usage was recorded

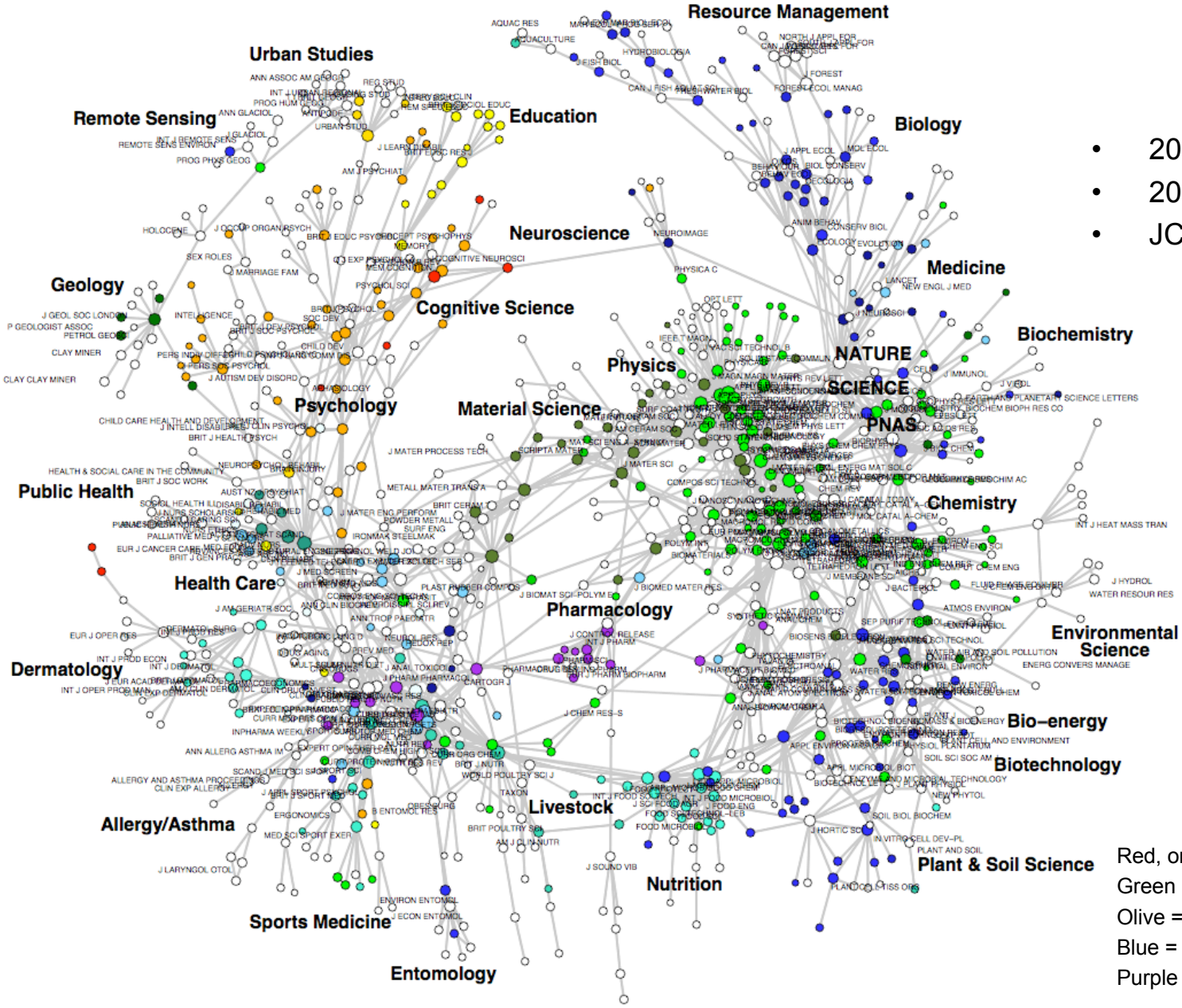
Note: not something we invented

- Association rule learning in data mining
- Cf. Netflix, Amazon recommendations

MESUR: Project Phases

1) Usage data acquisition
2) Science mapping from usage graphs
3) Metrics survey
4) Services





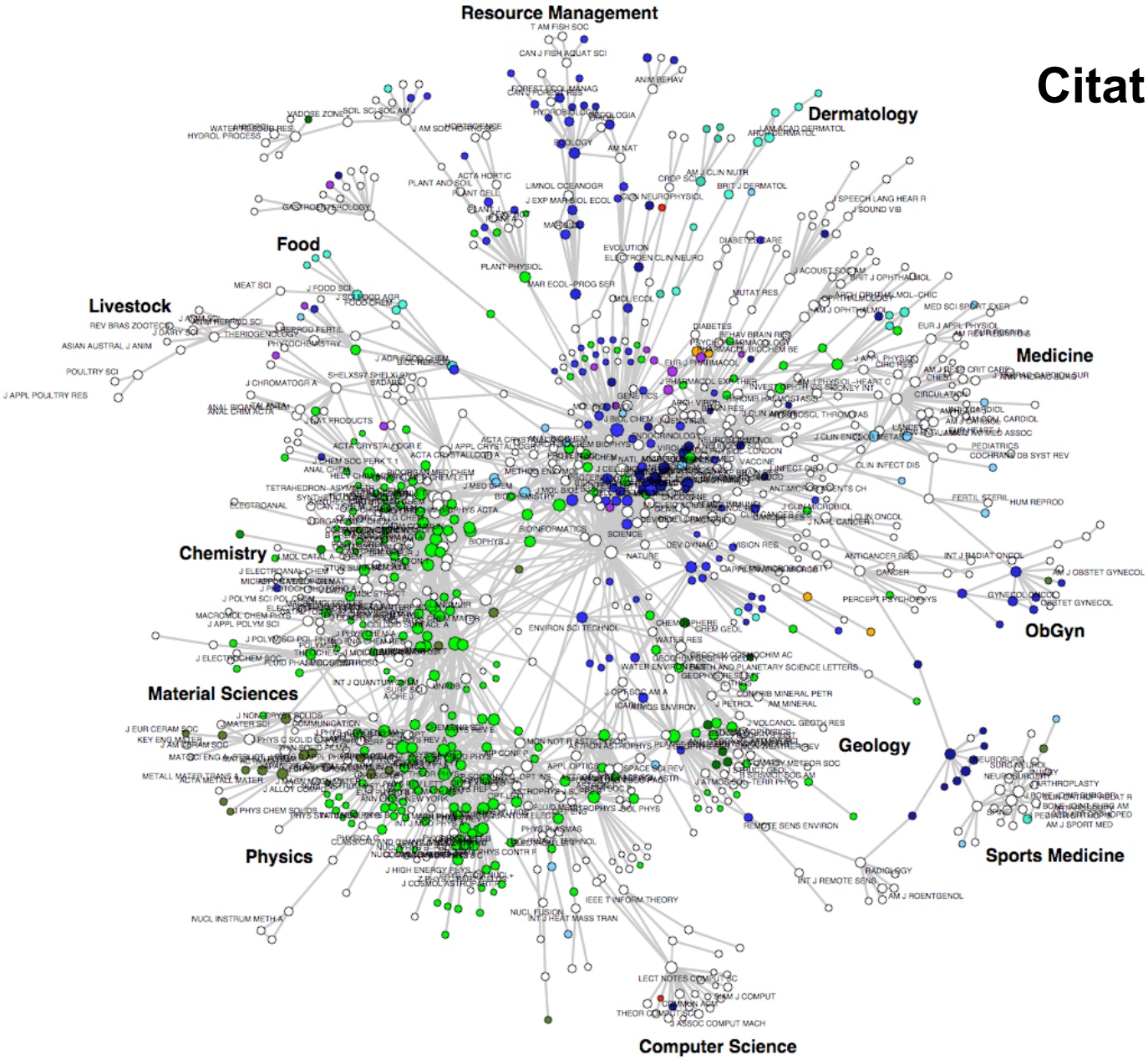
Usage map

- 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
UKSG, Torquay, UK, April 7-9 2008

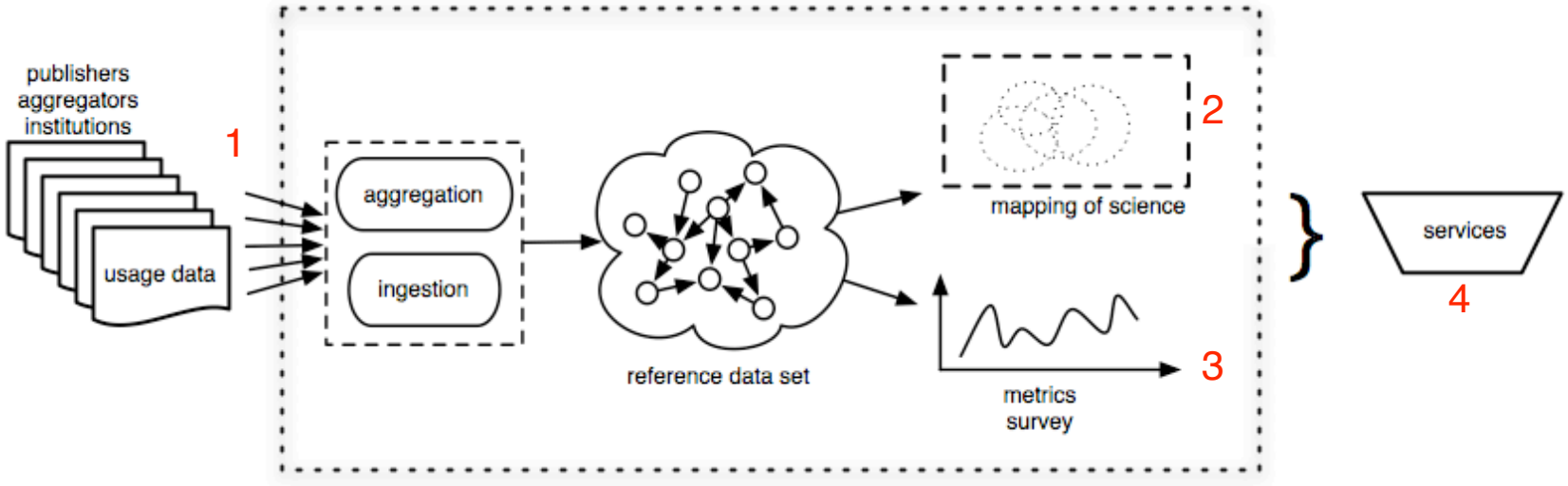
Citation map



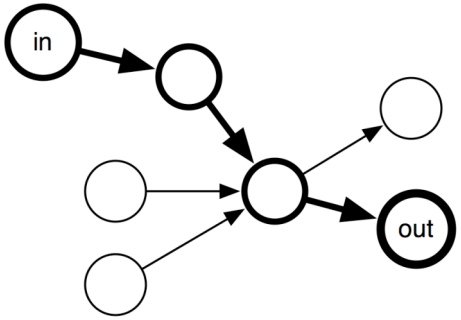
Digital Library Research & Prototyping Team
Research Library, Los Alamos National Laboratory
UKSG, Torquay, UK, April 7-9 2008

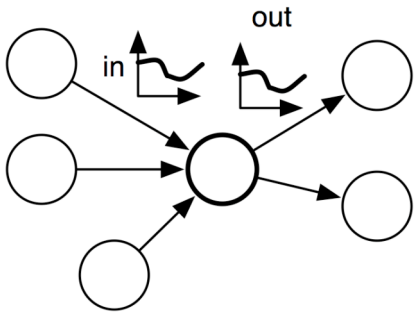
MESUR: Project Phases

1) Usage data acquisition
2) Science mapping from usage graphs
3) Metrics survey
4) Services



Network Metrics Computed for Usage and Citation Networks



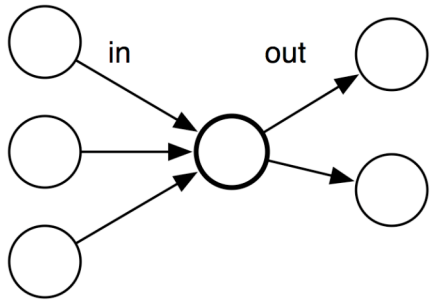


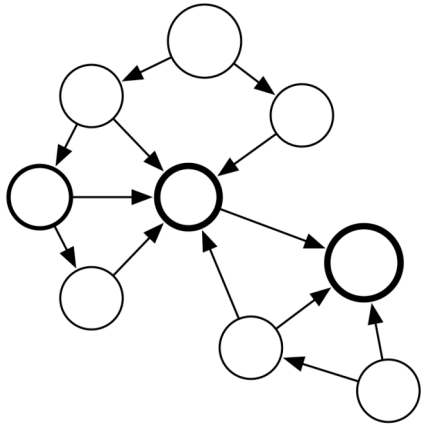
Citation-based metrics (JCR 2004)

- CITE-BE
- CITE-ID
- CITE-IE
- CITE-IF
- CITE-OD
- CITE-OE
- CITE-PG
- CITE-UBW
- CITE-UBW-UN
- CITE-UCL
- CITE-UCL-UN
- CITE-UNM
- CITE-UNM-UN
- CITE-UPG
- CITE-UPR
- CITE-WBW
- CITE-WBW-UN
- CITE-WCL
- CITE-WCL-UN
- CITE-WID
- CITE-WNM
- CITE-WNM-UN
- CITE-WOD
- CITE-WPR

Usage-based metrics: (MESUR 2006)

- USES-BE,
- USES-ID
- USES-IE
- USES-OD
- USES-OE
- USES-PG
- USES-UBW
- USES-UBW-UN
- USES-UCL
- USES-UCL-UN
- USES-UNM
- USES-UNM-UN
- USES-UPG
- USES-UPR
- USES-WBW
- USES-WBW-UN
- USES-WCL
- USES-WCL-UN
- USES-WID
- USES-WNM
- USES-WNM-UN
- USES-WOD
- USES-WPR





Citation 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

Citation Pagerank

	value	journal
1	0.0116	SCIENCE
2	0.0111	J BIOL CHEM
3	0.0108	NATURE
4	0.0101	PNAS
5	0.006	PHYS REV LETT

betweenness

	value	journal
1	0.076	PNAS
2	0.072	SCIENCE
3	0.059	NATURE
4	0.039	LECT NOTES COMPUT SC
5	0.017	LANCET

Closeness

	value	journal
1	7.02e-05	PNAS
2	6.72e-05	LECT NOTES COMPUT SC
3	6.43e-05	NATURE
4	6.37e-05	SCIENCE
5	6.37e-05	J BIOL CHEM

In-Degree

	value	journal
1	3448	SCIENCE
2	3182	NATURE
3	2913	PNAS
4	2190	LANCET
5	2160	NEW ENGL J MED

In-degree entropy

	Value	journal
1	9.849	LANCET
2	9.748	SCIENCE
3	9.701	NEW ENGL J MED
4	9.611	NATURE
5	9.526	JAMA

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

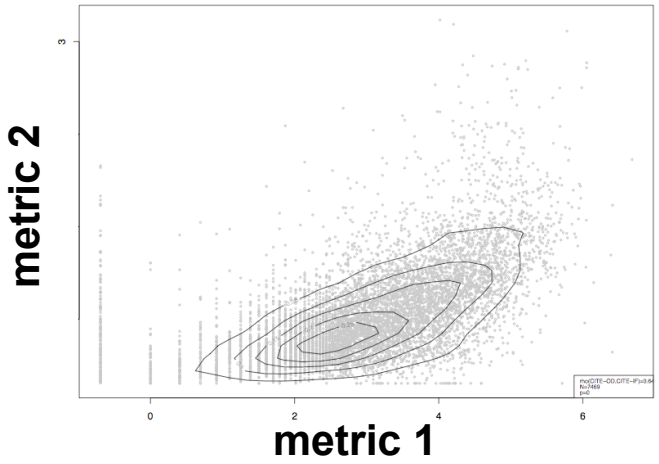
	value	journal
1	4195	SCIENCE
2	4019	NATURE
3	3562	PNAS
4	2438	J BIOL CHEM
5	2432	LNCS

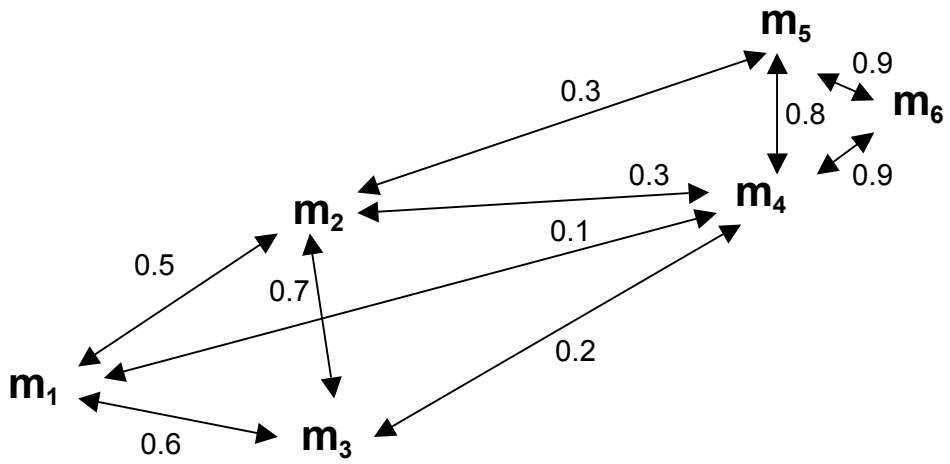
In-degree entropy

	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

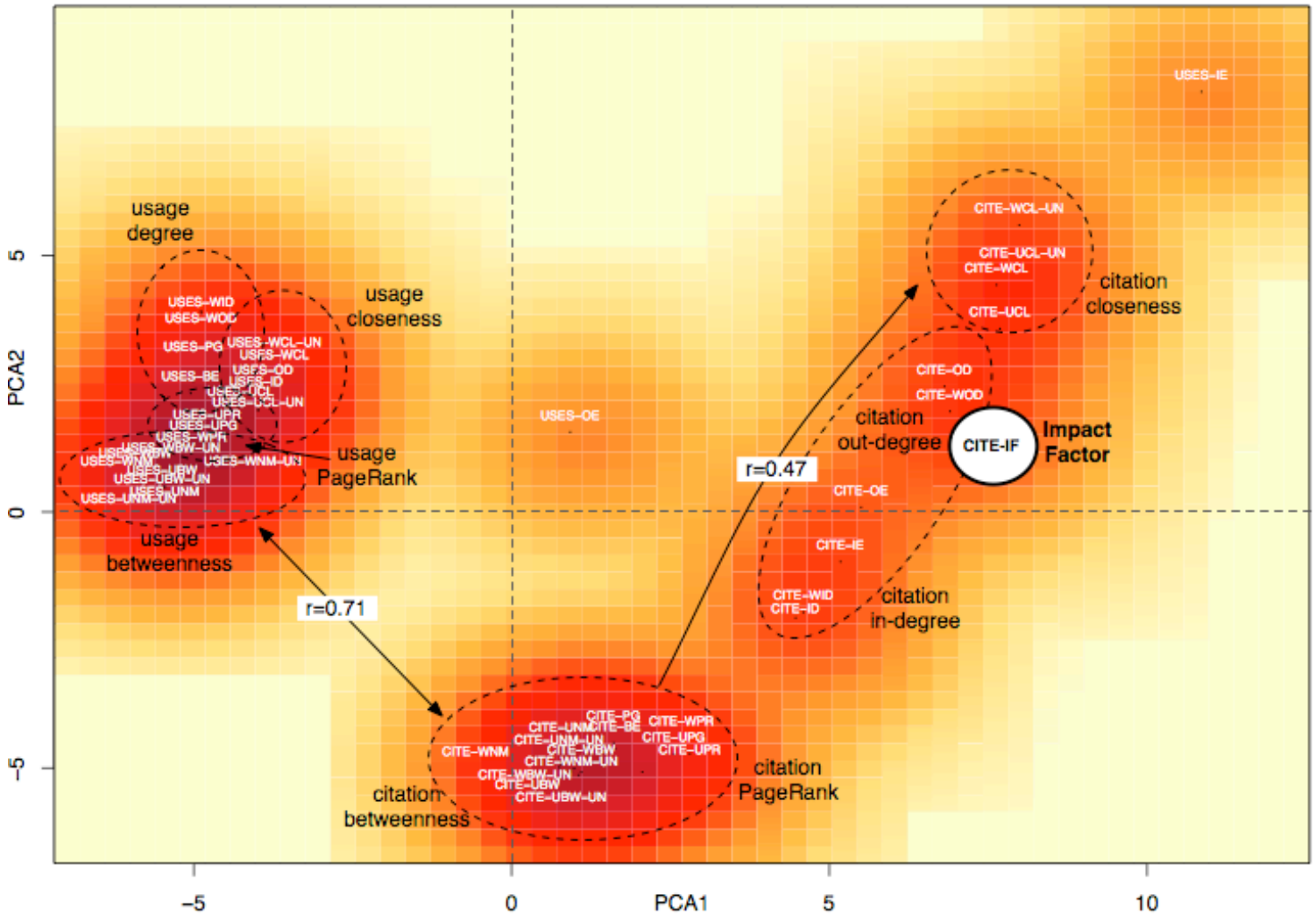
Metric Correlations: Metric Maps

	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10
m1	1.00	0.75	0.67	0.61	0.46	0.57	0.99	0.79	0.79	0.40
m2	0.75	1.00	0.96	0.81	0.82	0.83	0.73	0.68	0.69	0.77
m3	0.67	0.96	1.00	0.77	0.77	0.81	0.65	0.62	0.63	0.72
m4	0.61	0.81	0.77	1.00	0.64	0.67	0.60	0.50	0.51	0.64
m5	0.46	0.82	0.77	0.64	1.00	0.92	0.44	0.57	0.58	0.89
m6	0.57	0.83	0.81	0.67	0.92	1.00	0.55	0.65	0.66	0.77
m7	0.99	0.73	0.65	0.60	0.44	0.55	1.00	0.78	0.79	0.39
m8	0.79	0.68	0.62	0.50	0.57	0.65	0.78	1.00	0.99	0.54
m9	0.79	0.69	0.63	0.51	0.58	0.66	0.79	0.99	1.00	0.55
m10	0.40	0.77	0.72	0.64	0.89	0.77	0.39	0.54	0.55	1.00



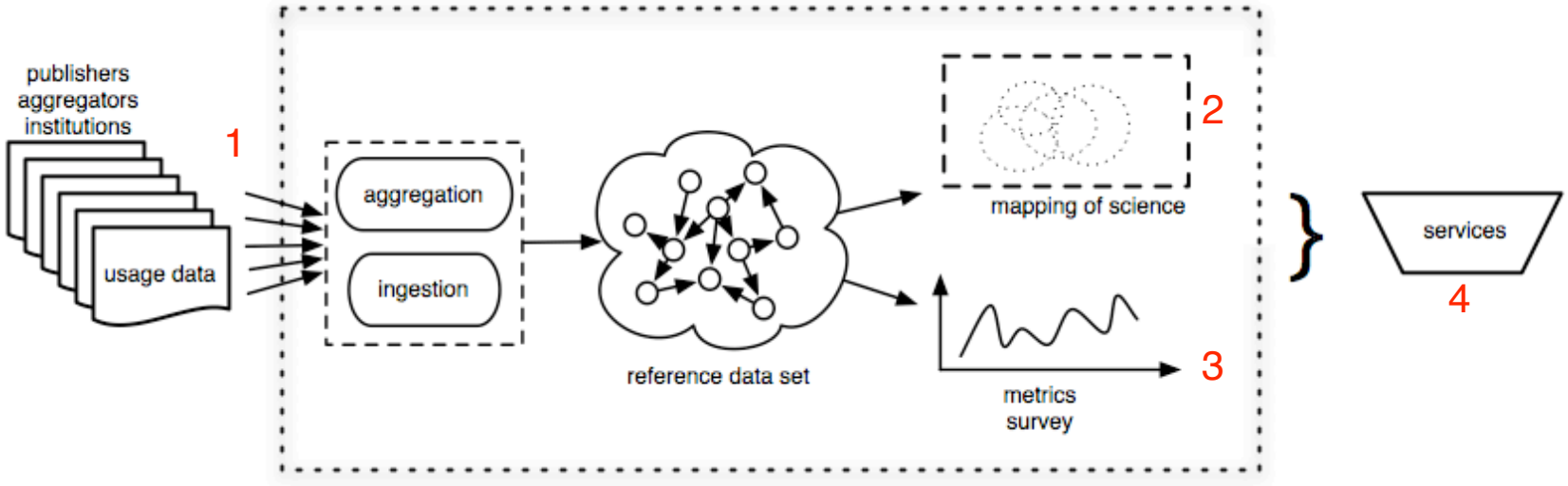


The MESUR Metrics Maps



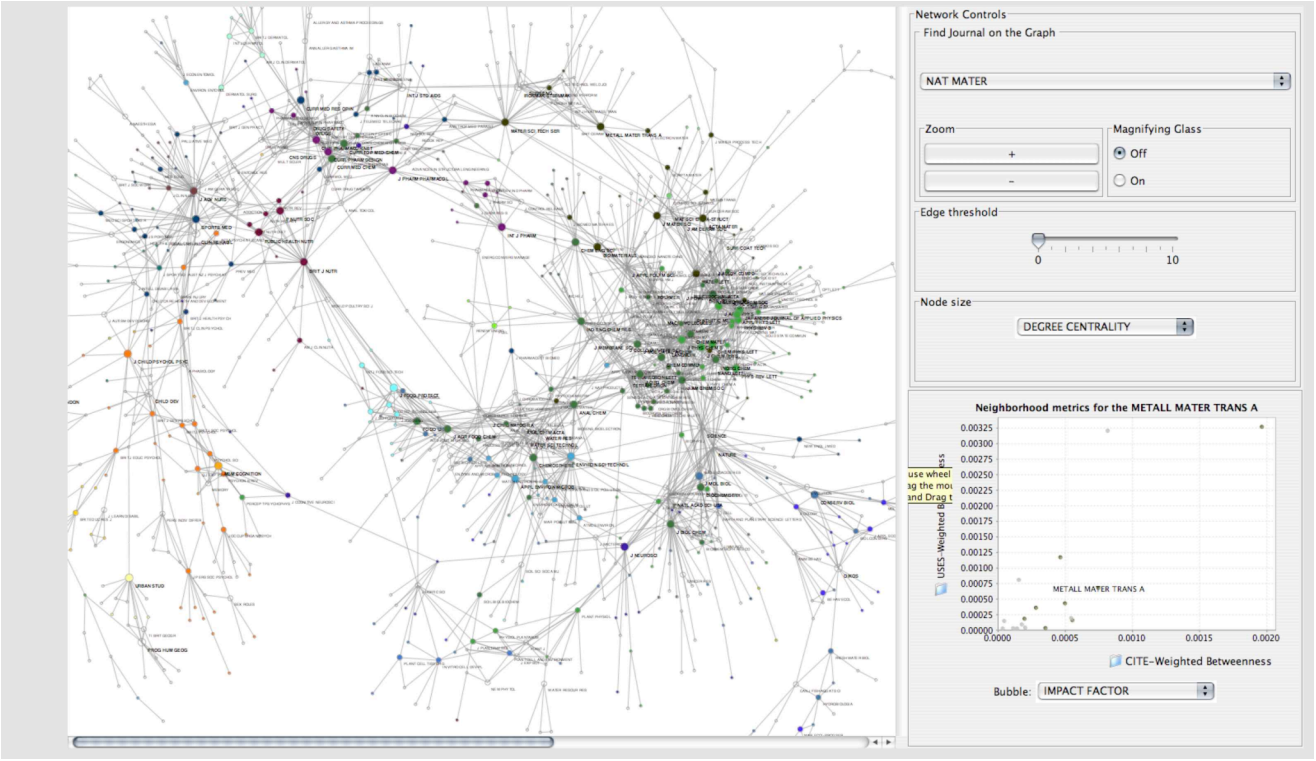
MESUR: Project Phases

1) Usage data acquisition
2) Science mapping from usage graphs
3) Metrics survey
4) Services



MESUR Explorer Prototype

- Based on MESUR usage data collection
- Explore large-scale usage maps of science
- Explore journal rankings according to multiple metrics of interest



Conclusions, 18 Months into MESUR

Usage data totally rocks!

- First scientific exploration of new paradigm for scholarly assessment
- Creation of a vast reference data set of usage data
- Infrastructure for a continued research program
- Beyond discussion of merits and validity of usage data

Even though many challenges remain. Obviously.

Publications related to MESUR

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.