MESUR:
metrics from scholarly usage of resources

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), Lyudmila L. Balakireva (LANL)
Wenzhong Zhao (LANL), Aric Hagberg (LANL)

MESUR is supported by the Andrew W. Mellon Foundation.

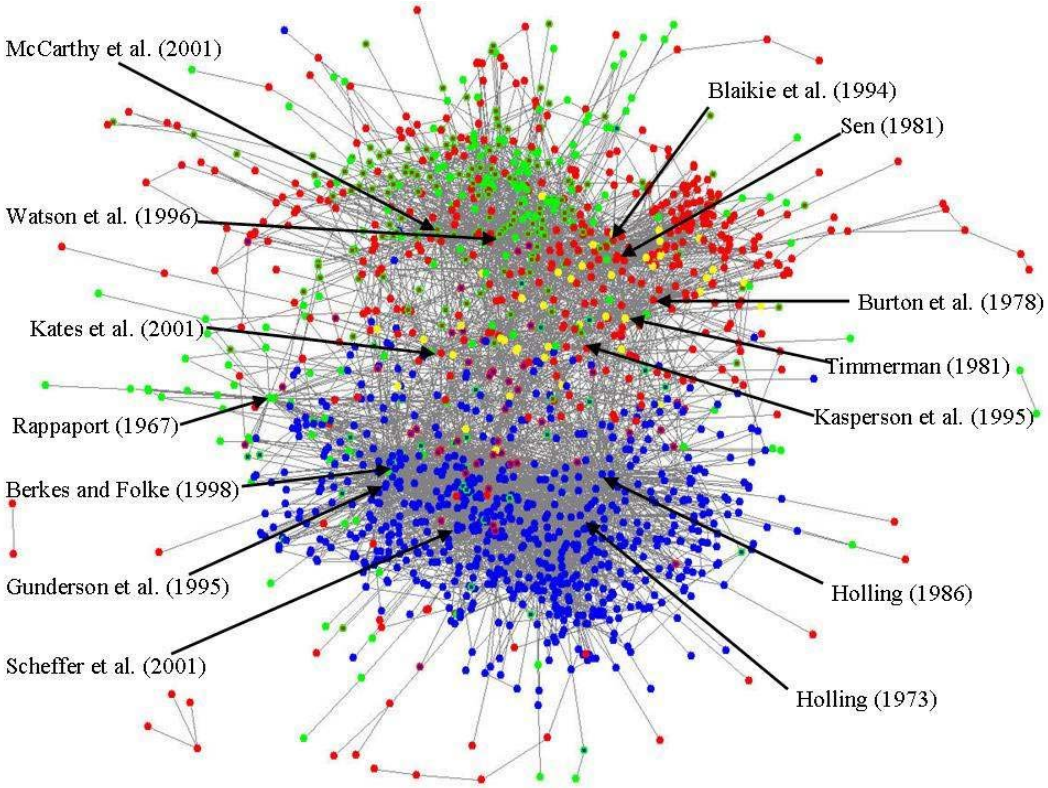
The MESUR team.

Johan Bollen (LANL): Principal investigator.
Herbert Van de Sompel (LANL): Architectural consultant.
Aric Hagberg (LANL): Mathematical and statistical consultant.
Marko Rodriguez (LANL): PhD student (Cognitive Science, UCSC).
Lyudmila Balakireva (LANL): Database management and development.
Wenzhong Zhao (LANL): Data processing, normalization and ingestion.

"The project's 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."

http:/www.mesur.org/

Scholarly community facts.



Janssen, M. A. (2006). Scholarly networks… in Global Environmental Change 16(3)

+- 1,510M articles indexed (WoS 2005)*
+- 1,230M ISSNs registered (2006)*
+- 1B full-text downloads (Elsevier Science Direct)*
+- 6.5B humans.

Similar predicament to web:
- Too much to track.
- Too many relationships.
- Even greater need to evaluate what really matters.

* Severe underestimates!

Scholarly evaluation matters.

Zero-sum world requires proper allocation of resources:

- Appointment decisions
- Funding decisions
- Monitoring trends and identify emerging focii
- Prioritize activities and attention

Evaluation of scholarly status

- Qualitative methods:
  - Tenure committees
  - Peer review
  - Policy-maker decisions
  - Personal judgment and experience

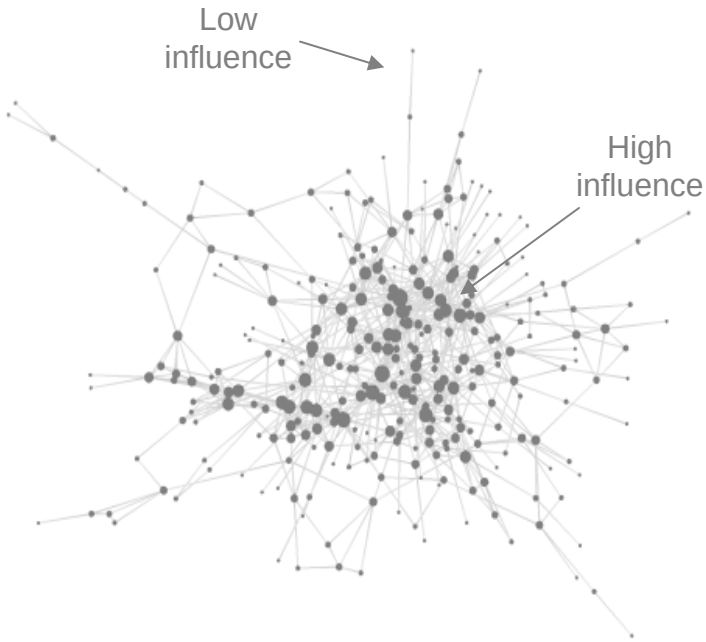
But scale is problem, thus:

- Quantitative indicators:
  - Performance metrics
  - Quality and status metrics

Citations and the journal impact factor.

Citation data:

- Golden standard of scholarly evaluation
- Citation = scholarly influences.
- Extracted from published materials.
- Main bibliometric data source for scholarly evaluation.

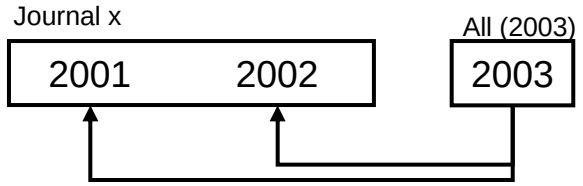


Journal Impact Factor: mean 2-year citation rate

2003 citations to 2001 and 2002 articles in X

divided by

number of articles published in X in 2001 and 2002



Widely applied

- Fair approximation of journal "status",…but
- Used to rank authors, departments, institutions, regions, nations, etc.
- Now common in tenure, promotion and other evaluation procedures!

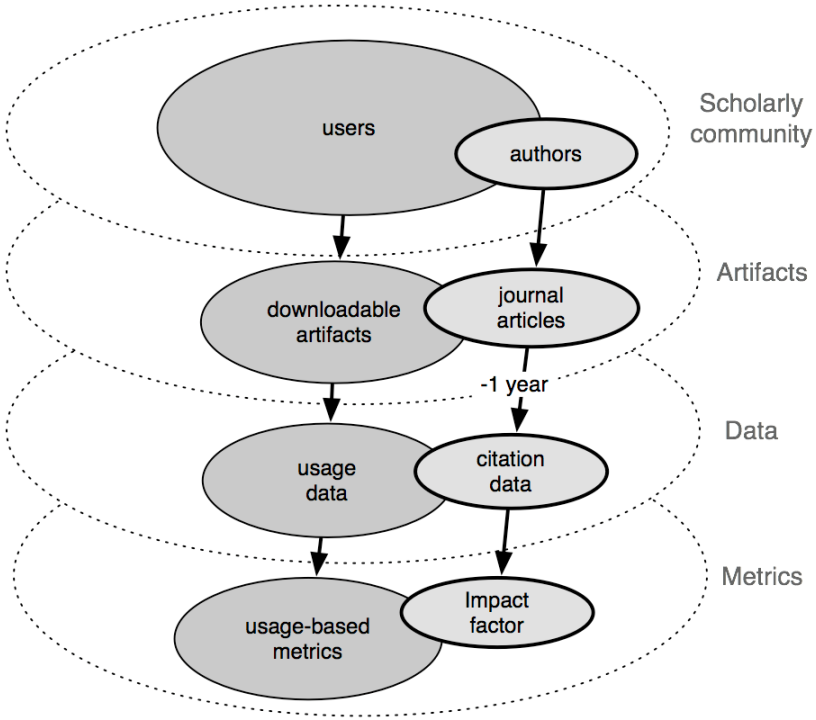
Can we improve on citation data and the impact factor?

Citation data pertains to 4 levels in the scholarly communication process:

- Community: authors of journal articles.
- Artifacts: journal articles.
- Data: citation data (+1 year publication delay).
- Metrics: mean citation rate rules supreme.
- Scale: expensive to extract.

However, for usage data:

- Community: all users including most authors.
- Artifacts: all that is accessible.
- Data: recorded upon publication.
- Metrics: a range of web and web2.0 inspired metrics, e.g. clickstream and datamining.
- Scale: automatically recorded at point of service.

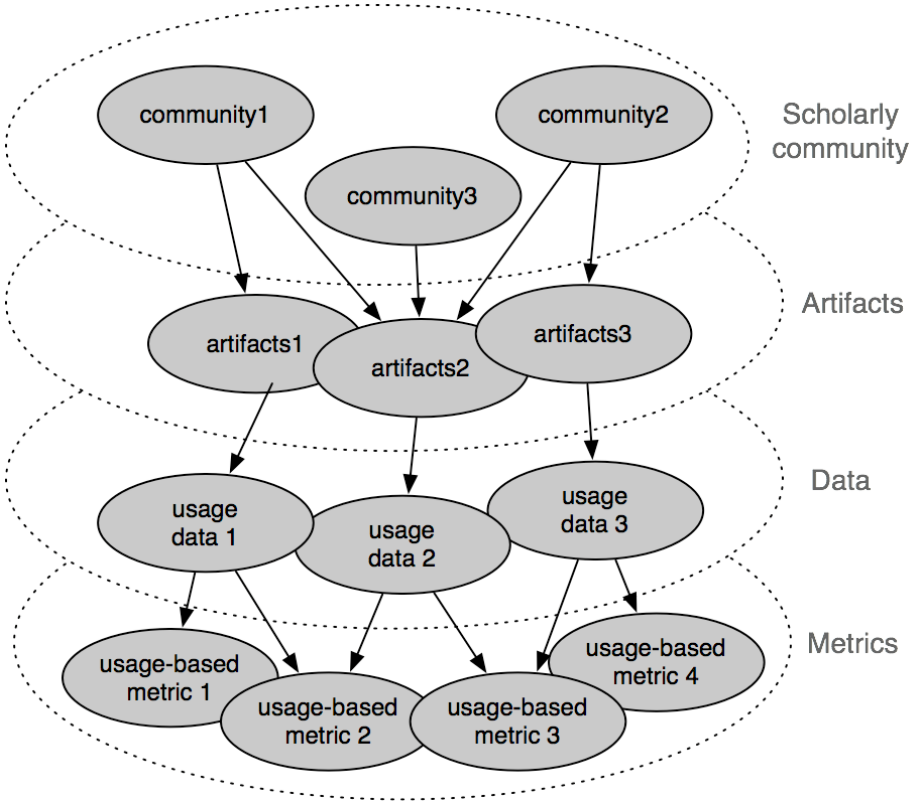


Hence, various initiatives focused on usage data: COUNTER, IRS, SUSHI, CiteBase. But where are the metrics?

Challenges to usage-based metrics.

Usage-based metrics have lagged development. Here's why:

- Community: particular digital services.
- Artifacts: institutional policies and subscriptions.
- Data: particular sub-communities and collections of artifacts.
- Metrics: variety of possible metrics. What do they mean?



Different outcomes for different communities.

	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

Los Alamos

	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

Cal. State

Solution: TREC model.

Create reference data set so that metrics can compete or be evaluated on same substrate.

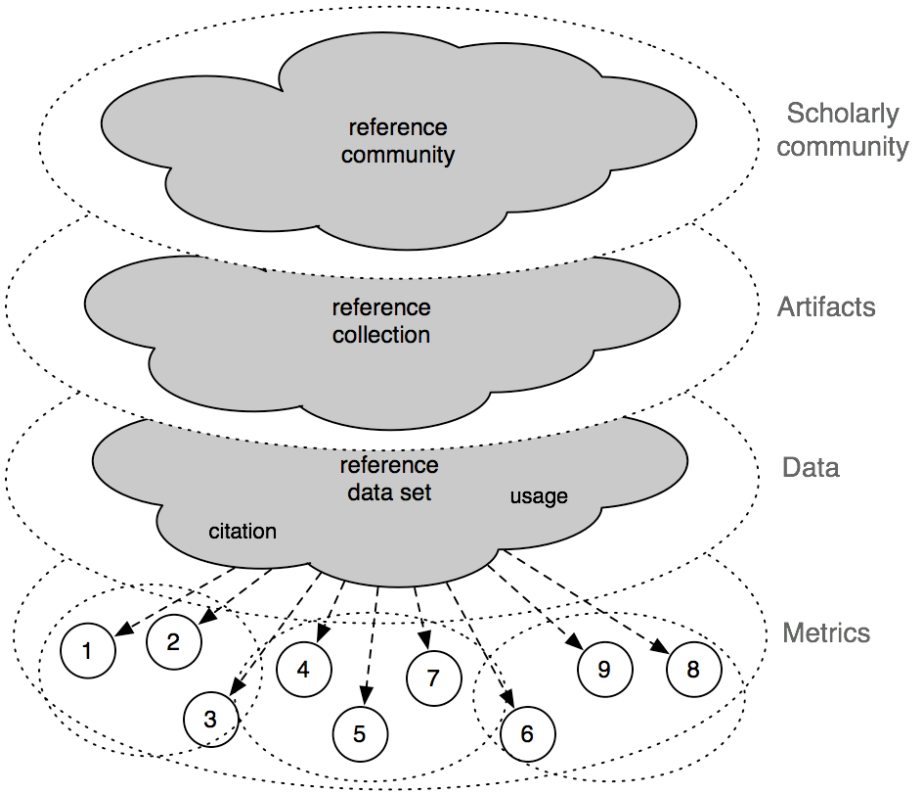
MESUR: Metrics from Scholarly Usage of Resources.

Andrew W. Mellon Foundation funded study of usage-based metrics (2006-2008)

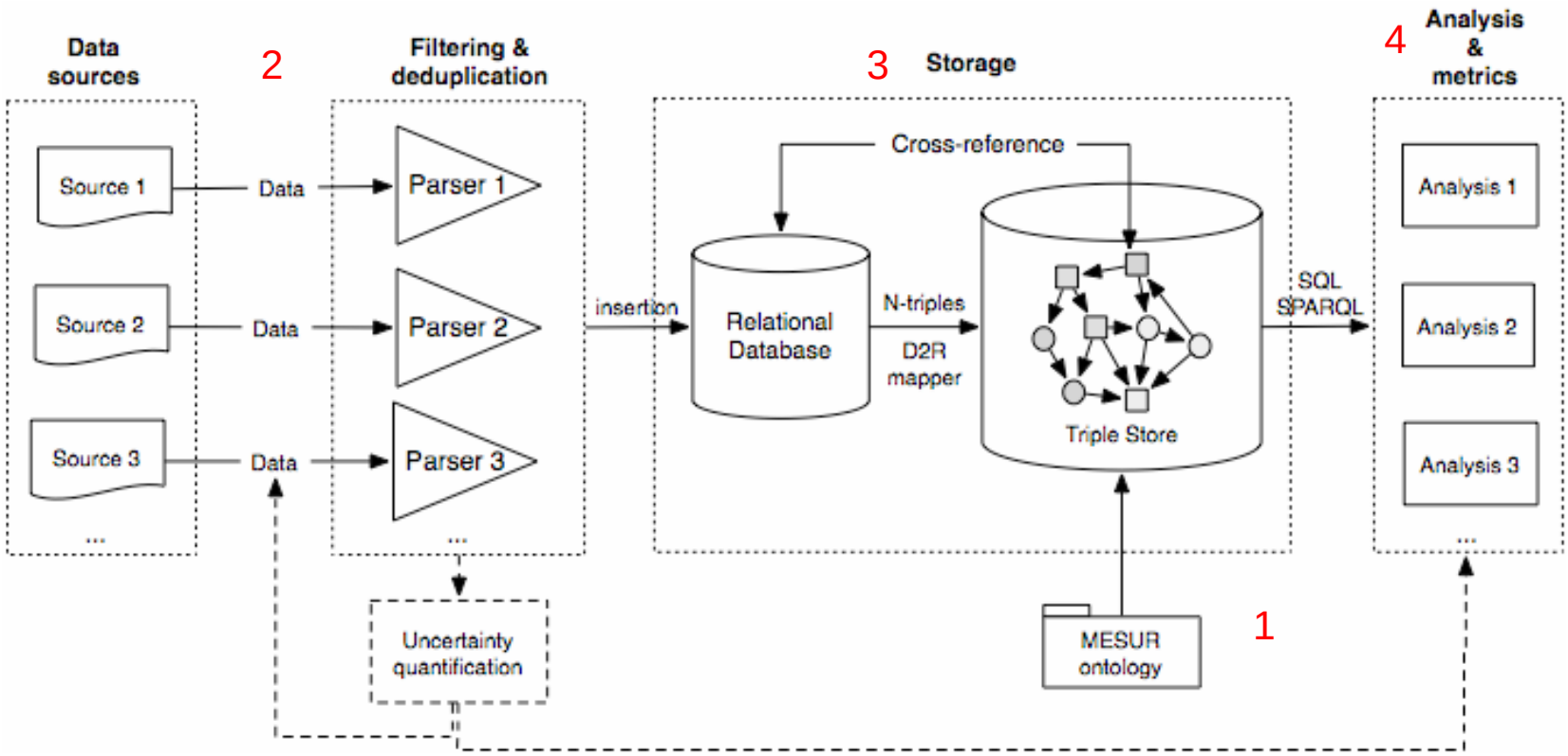
Executed at the Digital Library Research and Prototyping team, Los Alamos National Laboratory Research Library

Objectives:

1. Create a **model** of the scholarly communication process.
2. Create a large-scale reference data set (semantic network) that relates all relevant **bibliographic, citation and usage data** according to (1).
3. **Characterize** reference data set (2)
4. **Survey usage-based metrics** on basis of reference data set**.**

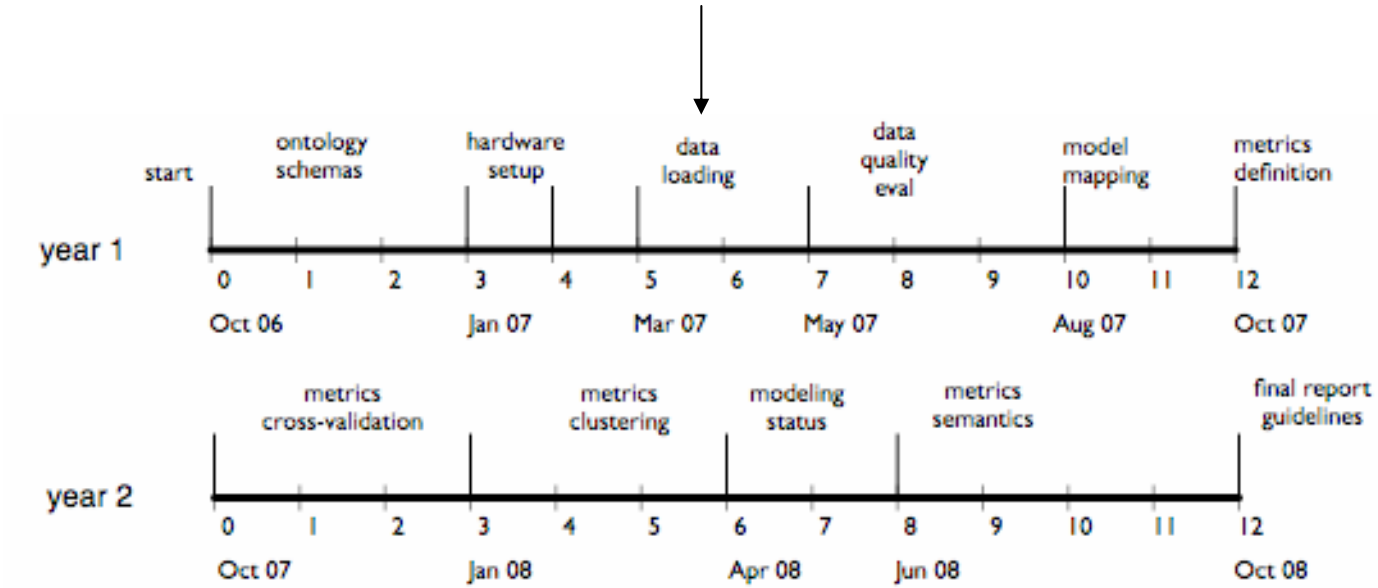


Project data flow and work plan.



Project timeline.

We are here!



Progress so far

Phase 1: Model of scholarly communication process:

- RDF/OWL ontology completed
- http://www.mesur.org/schemas/2007-01/mesur/
- Rodriguez et al. **A Practical Ontology for the Large-Scale Modeling of Scholarly Artifacts and their Usage,** JCDL07

Phase 2a. Agreements:

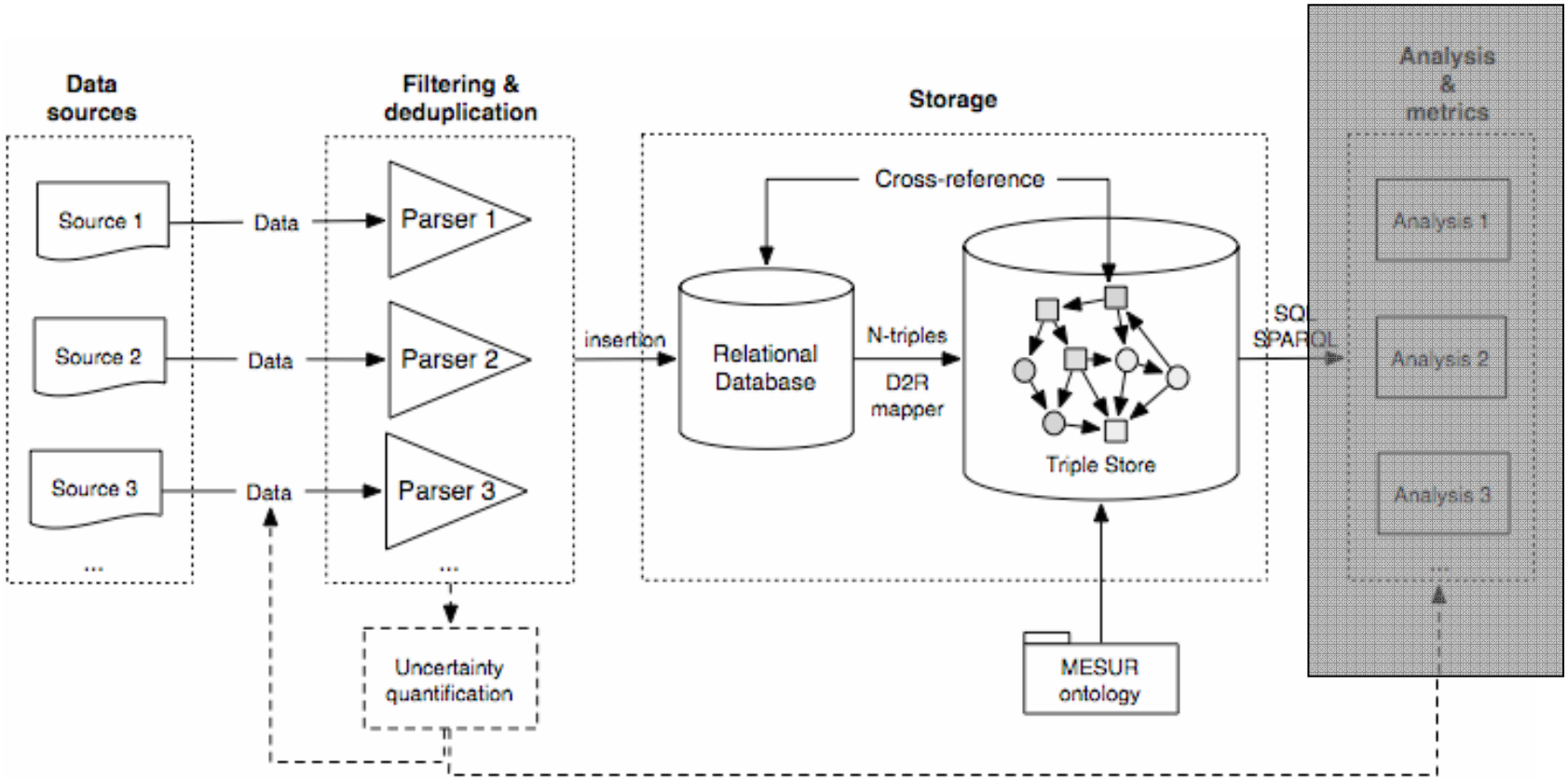
- 4 major university consortia
- 3 major aggregators
- 5 major publishers
- Support of COUNTER project
- More coming

Phase 2b. Data loaded and normalized:

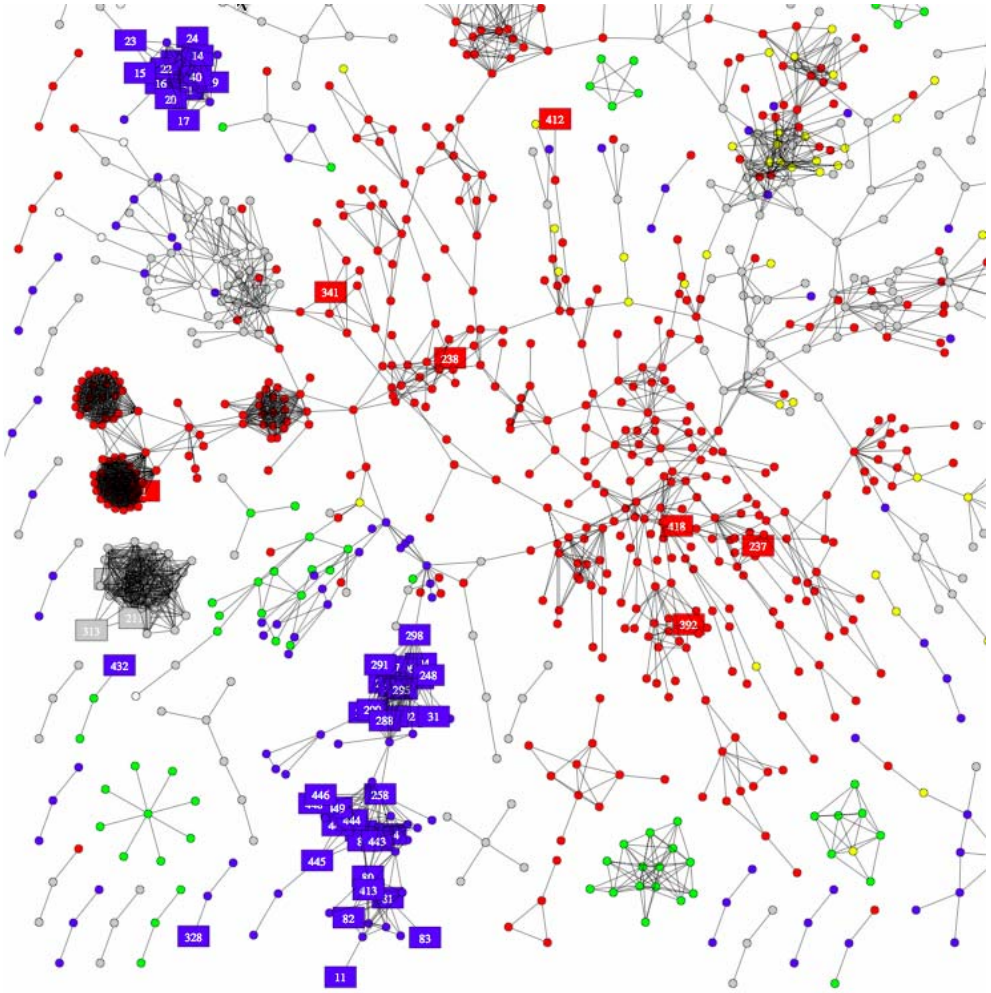
- Usage:
  - LANL: 400,000, +1 year
  - UT: 2,979,679, +- 1 year
  - BMC: 24,000,000, 2 year
  - anonymous: > 1,000,000 5-years
  - anonymous: > 2,500,000 1-year
  - anonymous: > 50,000,000 1-week
  - …
- Initial focus on journal level:
  - 50,000,000 citation relations
  - 300,000 groups
- Article level:
  - Span: nearly 10 years
  - 30M documents
  - 500M citation relations
  - 75M agents

Phase 2c. MESUR semantic network: >10B triples

Characterization and metrics survey.



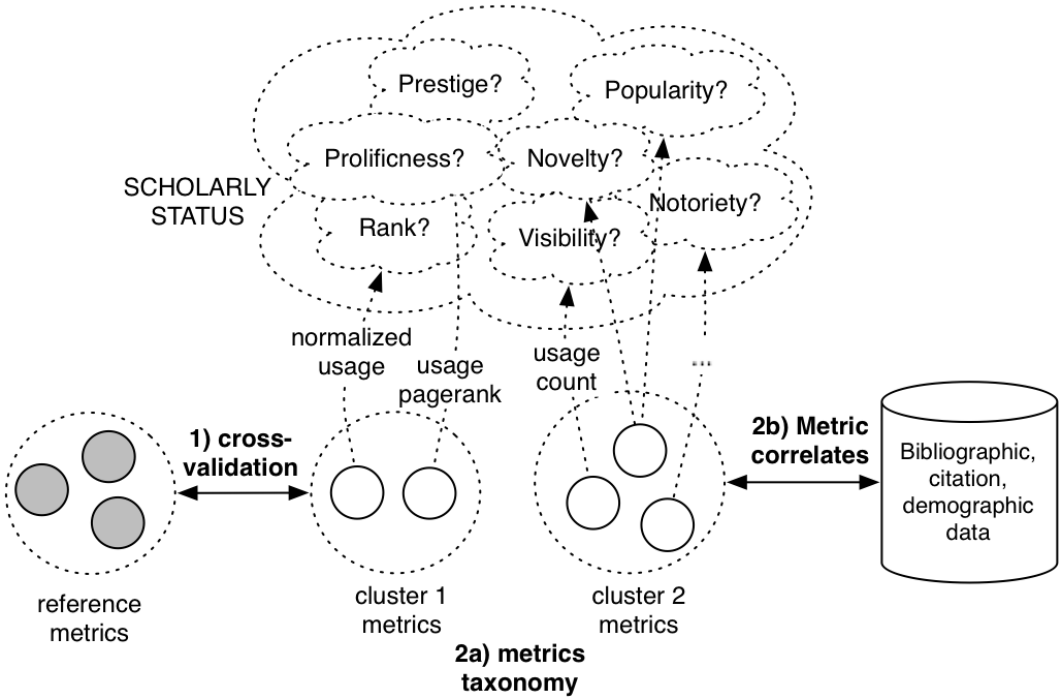
Phase 3: Semantic network characterization.



Objective is to study topological features of scholarly community and scholarly communication process similar to web characterization studies.

- Quantitative topology:
  - Social network topology
  - Small-world properties
  - Preferential attachment
- Qualitative mappings and visualization.
  - Spectral analysis
  - MDS
  - VxInsight
  - Longitudenal analysis
- Semantic features:
  - Cross-validation
  - Sub-networks

Phase 4. Metrics survey.



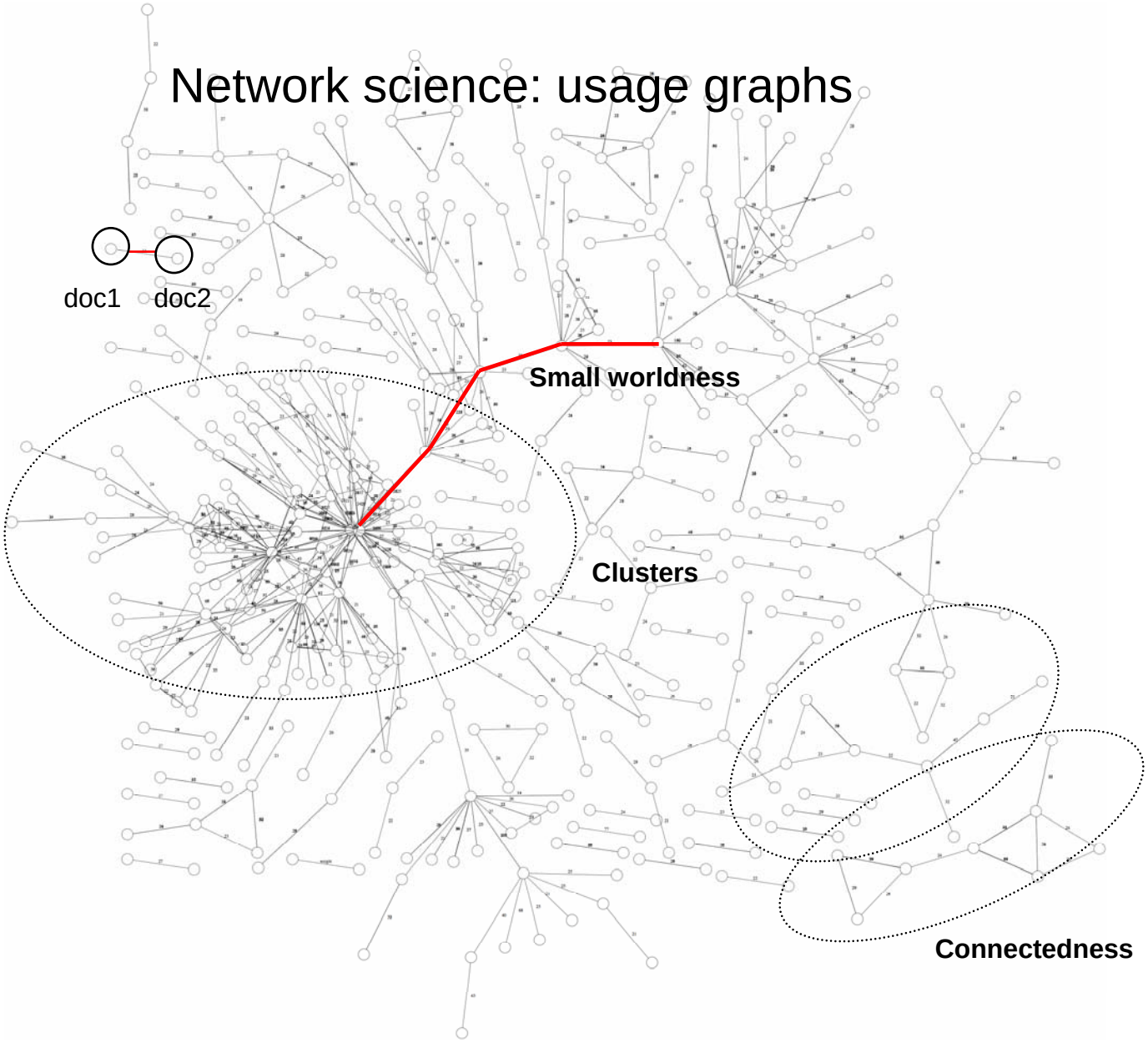
Three phased approach to perform survey:

1. Cross-validation
   - Set of reference metrics:
     - COUNTER usage statistics
     - Impact Factor
     - Citation counts
   - Determining correlations with variety of defined usage-based metrics
2. Metrics taxonomy:
   - **Hypothesis**: true metrics of scholarly impact all measure different yet related aspects of impact (cf. Big Five model)
   - Determining correlation structure of various metrics: facets of impact
   - Assessment of metric clusters
3. Metric correlates:
   - Use of bibliographic analysis (text and language processing)
   - Determining main correlates of defined metrics

Case study

- University of Texas Libraries at The University of Texas at Austin.
- Data characteristics:
  - 1.5 year of data, 2.9M usage events
  - SFX link resolver logs
  - Focus on fulltext downloads
- Combined with JCR citation data
  - 1998 to 2005, 50,926,947 citation relationships
  - Citation graph derived from JCR citation data (8,000 journals)
  - IFs and number of articles included for all years
- Usage graph:
  - 42,595 number of journals
  - 436,871 number of edges
- Metrics calculated:
  - Citation: IF, Citation PageRank
  - Usage: Usage Factor, Usage Impact Factor 06, Usage PageRank, Weighted In-degree, Weighted Out-degree, In-Degree entropy, Out-Degree entropy

Network science: usage graphs



U. Texas rankings

IF (2005)	TITLE (ABBRV)
49.794	CA A CANCER
47.400	ANN REV IMMUNOL
44.016	NEJM
33.456	ANN REV BIOCHEM
31.694	NAT REV CANCER
30.458	NAT REV IMMUNOL
30.254	REV MOD PHYS
29.852	NAT REV MOL CELL BIOL
29.431	CELL
29.273	NATURE

CPR05	IF	TITLE (ABBRV)
0.010	5.854	J BIOL CHEM
0.009	29.273	NATURE
0.009	10.231	PNAS
0.006	7.489	PHYS REV LETT
0.005	44.016	NEJM
0.004	7.419	JACS
0.004	23.332	JAMA
0.004	2.784	J GEOPHYS RES
0.003	7.506	J NEUROSCIENCE
0.003	1.131	BLOOD

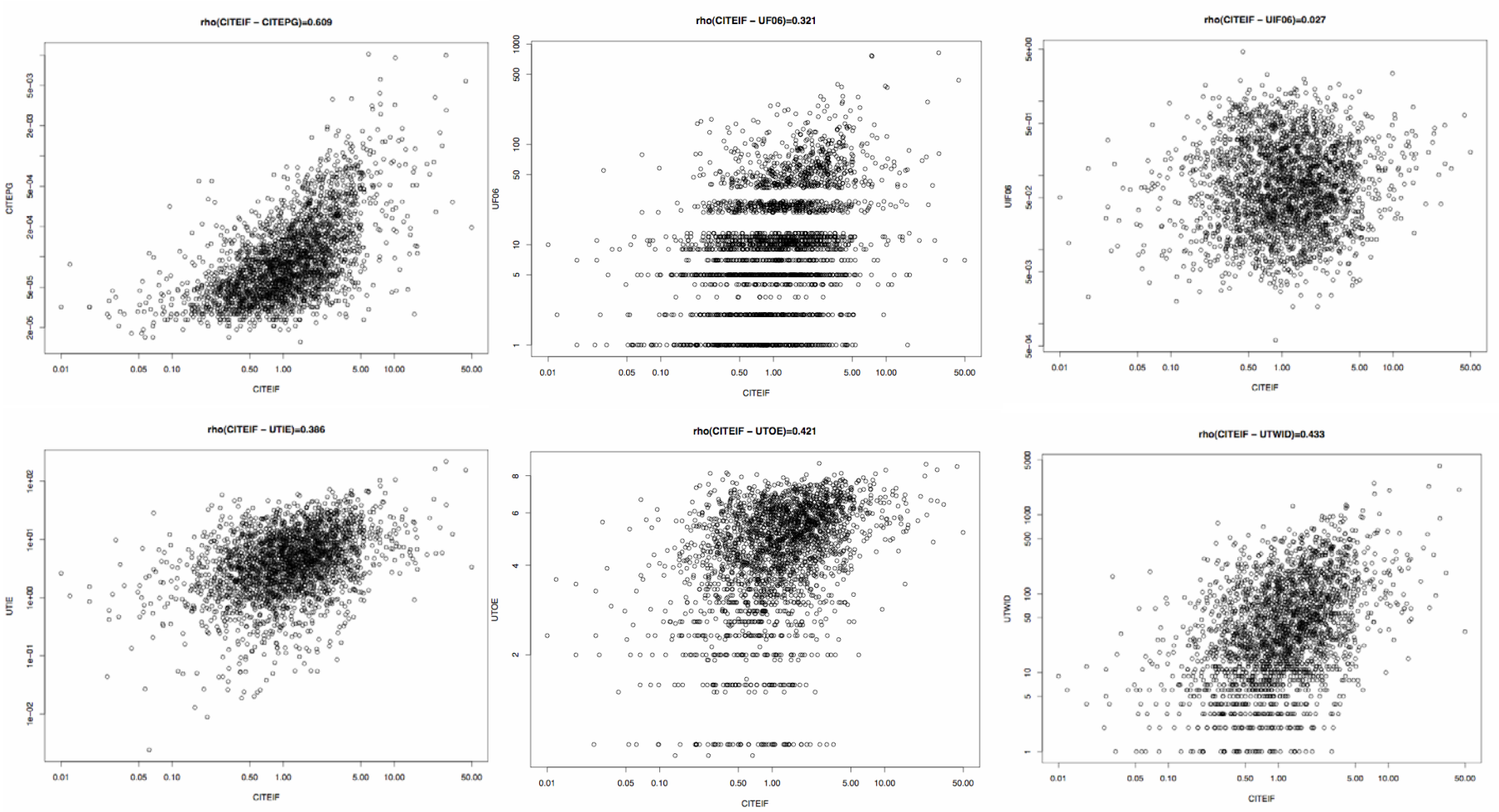
UIF06	IF	TITLE (ABBRV)
4.611	0.444	CONTEMP SOCIOL
2.346	9.885	BEHAV & BRAIN SCI
2.264	0.735	J MUSIC THERAPY
1.980	1.578	SOCIAL FORCES
1.889	0.422	SOCIOL SPECTRUM
1.736	1.623	AM HIST REV
1.672	3.262	AM J SOCIOL
1.632	0.204	WOMEN AND THERAPY
1.626	0.664	CHILD & YOUTH SERV REV
1.522	1.957	THE FUTURE OF CHILDREN

UPR05	IF	TITLE (ABBRV)
0.0034	29.273	NATURE
0.0022	23.332	JAMA
0.0021	44.016	NEJM
0.0017	7.419	JACS
0.0016	10.231	PNAS
0.0012	7.489	PHYS REV LETT
0.0011	1.350	J MARR & FAM
0.0011	4.113	J AM ACAD CHILD ADOLESC
0.0010	4.211	J PERS SOC PSYCH
0.0009	2.784	J OF GEOPHYS RES

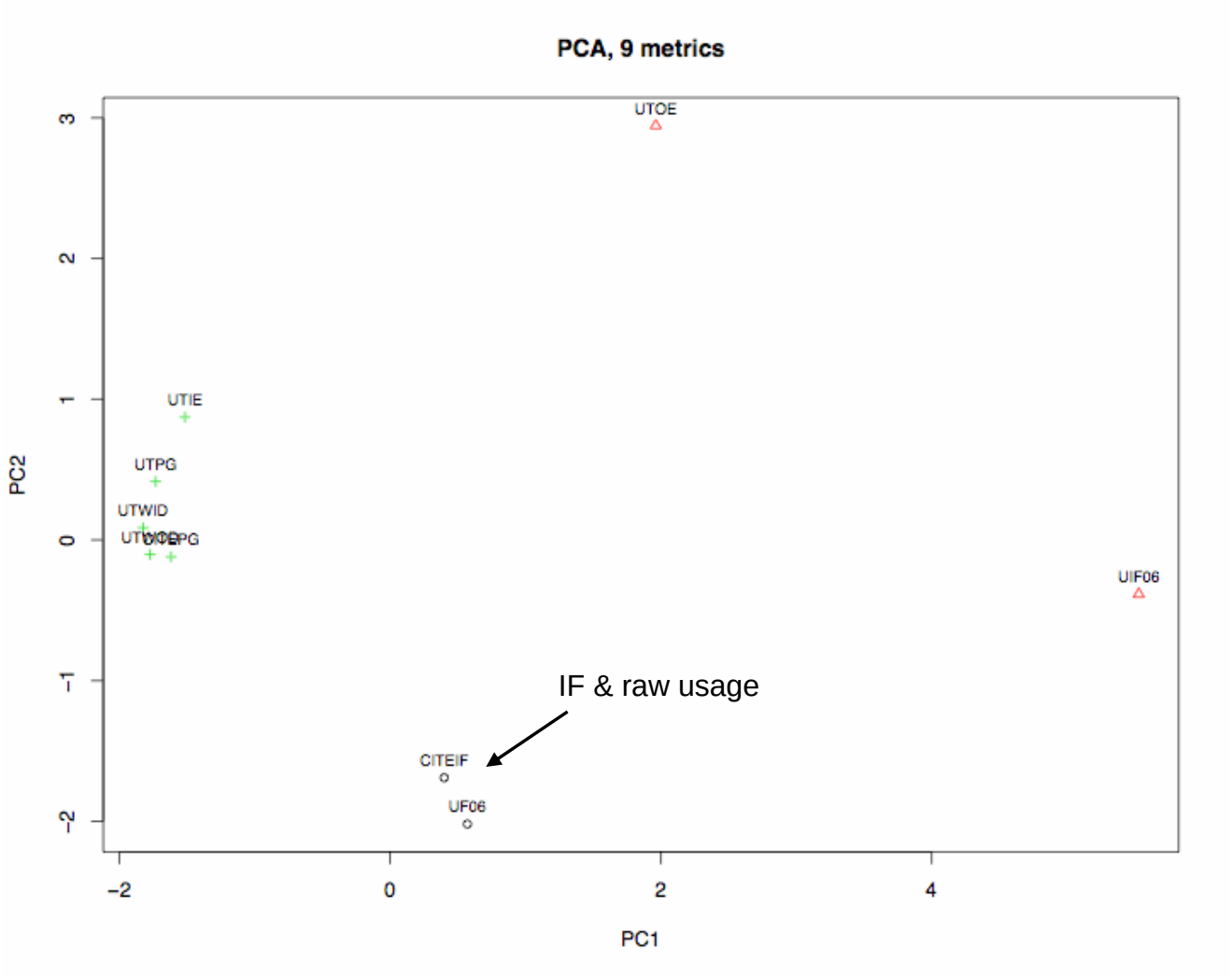
IDE	IF	TITLE (ABBRV)
217.103	29.273	NATURE
162.039	23.332	JAMA
155.234	44.016	NEJM
104.73	10.231	PNAS
102.557	7.419	JACS
82.246	7.489	PHYS REV LETT
71.752	2.619	SOC SCI & MED
68.707	4.113	J AM ACAD CHILD ADOLESC
66.982	5.635	ANAL CHEM
66.482	5.853	AM J CLIN NUTRITION

UIF06	IF	TITLE (ABBRV)
0.006	0.875	NUCL PHYS B
0.006	5.944	J HIGH ENERG PHYS
0.008	0.579	J MATH ANAL APPL
0.009	4.852	PHYS REV D
0.0010	0.777	ACTA CRYSTALLOGR C
0.0012	0.459	J ALGEBRA
0.0013	0.59	LINEAR ALGEBRA APPL
0.0013	0.581	ACTA CRYSTALLOGR E
0.0014	0.569	J COMPUT APPL MATH
0.0014	0.469	CR MATH

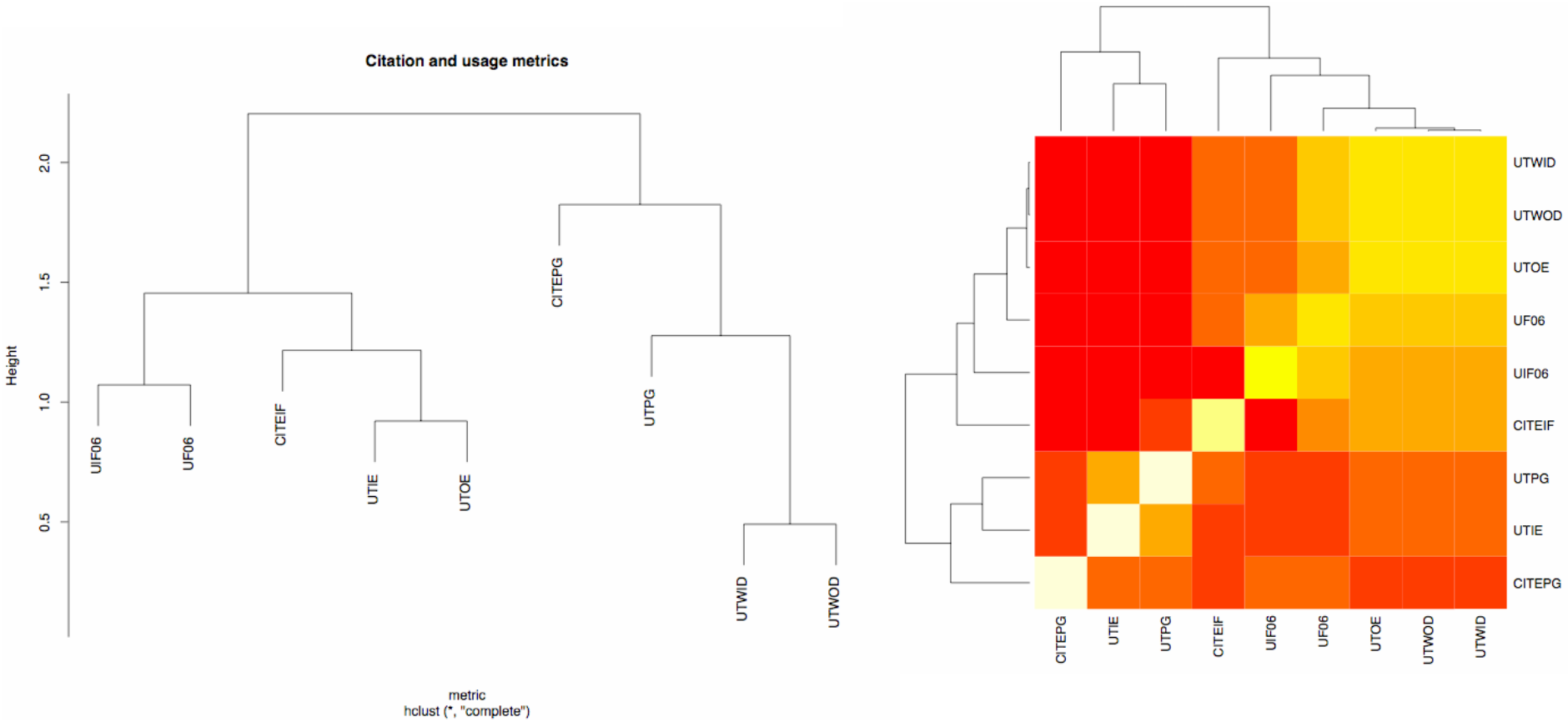
Quantitative comparisons



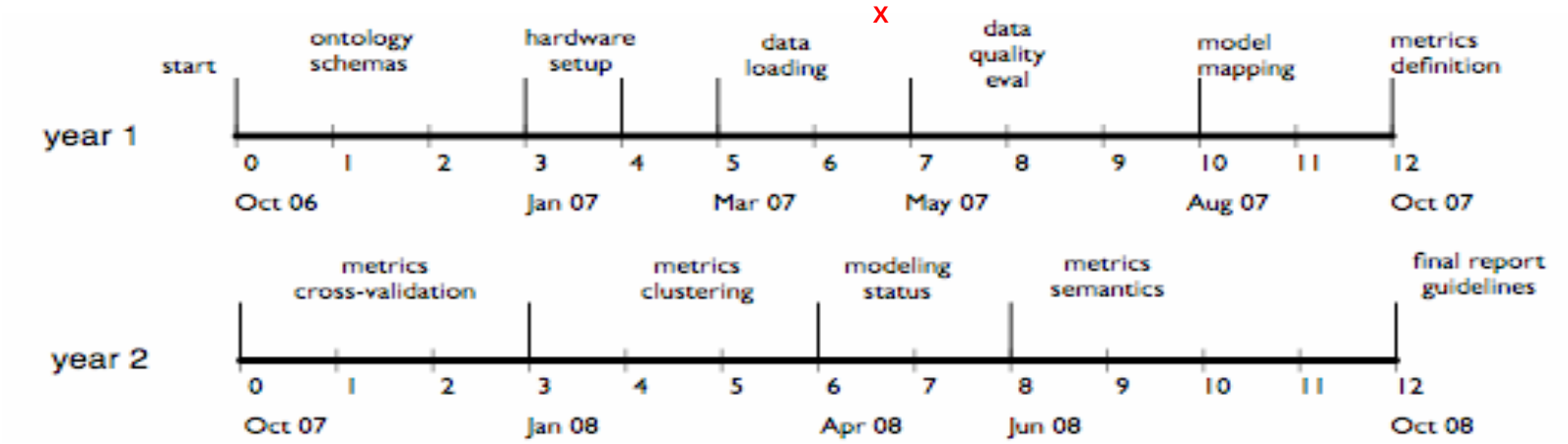
Metrics relationships



Metrics relationships



To be continued…



- **Now:** agreements with significant data-providers in place. Usage data ready to load.
- **Immediate future: data validation**
  - Loading journal-level data (even when derived from item-level data)
  - Execute small scale project deliverables on this data set
- **End of year: item level lay of the land**
  - Loading item-level data (in addition to journal-level)
  - Characterization of semantic networks, including usage graph
- **Next year: metrics definition**
  - Cross-validation to COUNTER statistics and IF
  - Correlation to bibliographic indicators.
  - Watch the MESUR web site for more data.

Issues and challenges

- 6 down, 18 to go
- Uncertainty quantification: methods to assess noise, errors and reliability of outcomes.
- Data:
  - Sampling problems: final data set representativeness?
  - Mixing aggregator, publisher and institutional data?
  - Creating reference data set: freeze in time or continue to develop?
- Scaling: analysis and transformation algorithms?
- Obtaining data: efforts continue
  - Publishers: reluctant, concerns over ownership and business implication
  - Aggregators: willingness to share, significant interest in outcome
  - Institutions: willingness to share , significant interest in outcome

Some relevant publications.

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

Johan Bollen, Herber Van d 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.