



# Measuring impact revisited

**Frank Scholze**  
**Stuttgart University Library**

**5th Workshop on Innovations in Scholarly Communication**  
**CERN, Geneva, 18.4.2007**



# Outline

1. Scope / definitions
2. Issues
3. Ongoing work / projects



## Impact

*“Impact is any change resulting from an activity, project, or organization. It includes intended as well as unintended effects, negative as well as positive, and long-term as well as short-term.”*

From: Susan Wainwright. Measuring Impact - A Guide to Resources. NCVO, 2002



# Measuring I

- What
  - Impact vs. output
- Publications are the quantifiable output of the research process
- Electronic publications as a limited but rather well defined sub-field of research output
- Which publications represent „e-science“?



## Measuring II

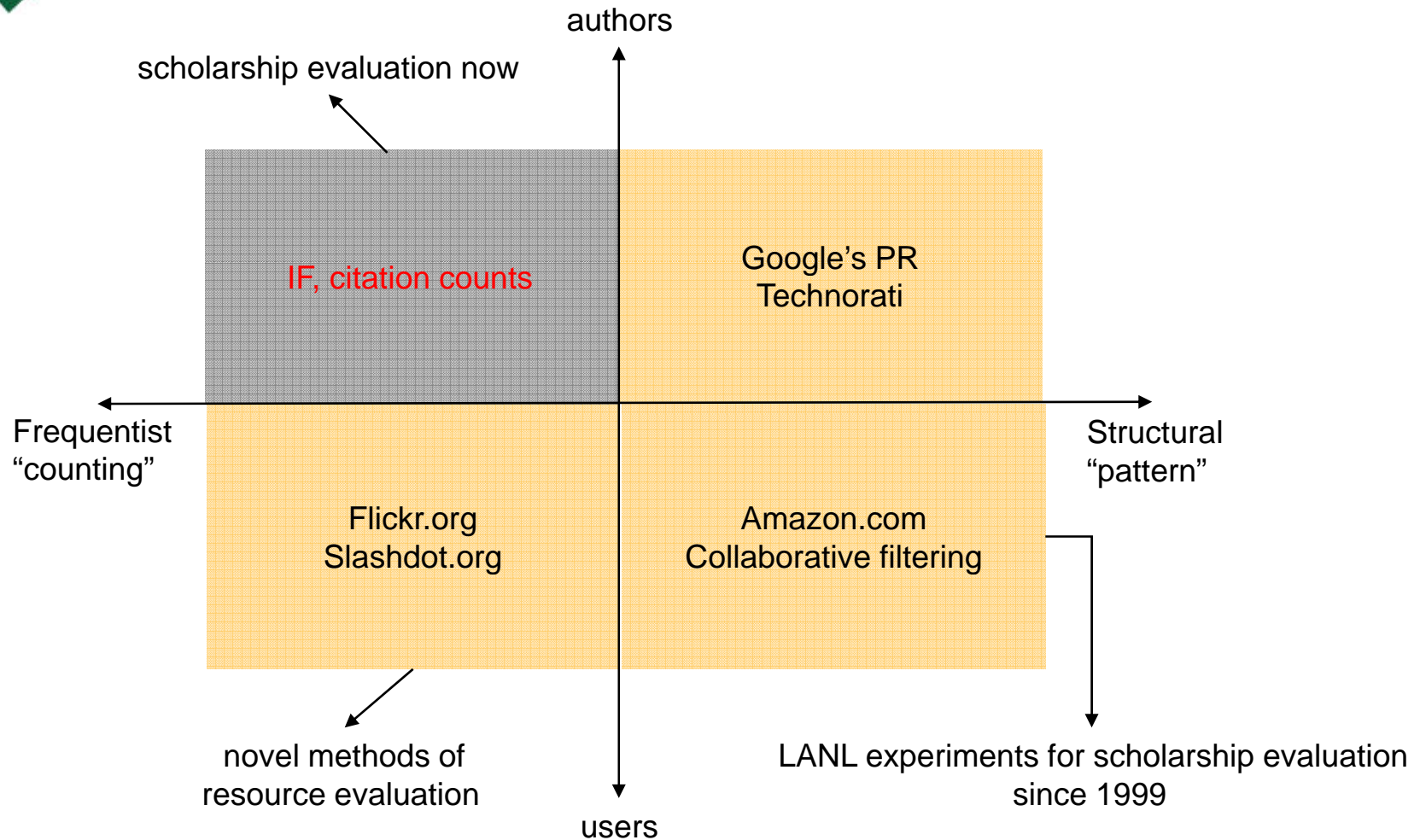
- How
  - Choose indicators for publications (citation, usage, „endorsement“, post-publication peer-review ...)
  - Collect information
  - Analyse / digest the information



# Indicators and methods

- Web Analytics = Usage
  - Logfile analysis
  - Server side scripts (e.g. Linkresolver)
  - Page tagging, web bugs, cookies
  - Hybrid methods
- Citation Analysis
  - Examination of the frequency and pattern of citations in publications
  - Impact factor (IF)

# A taxonomy of metrics

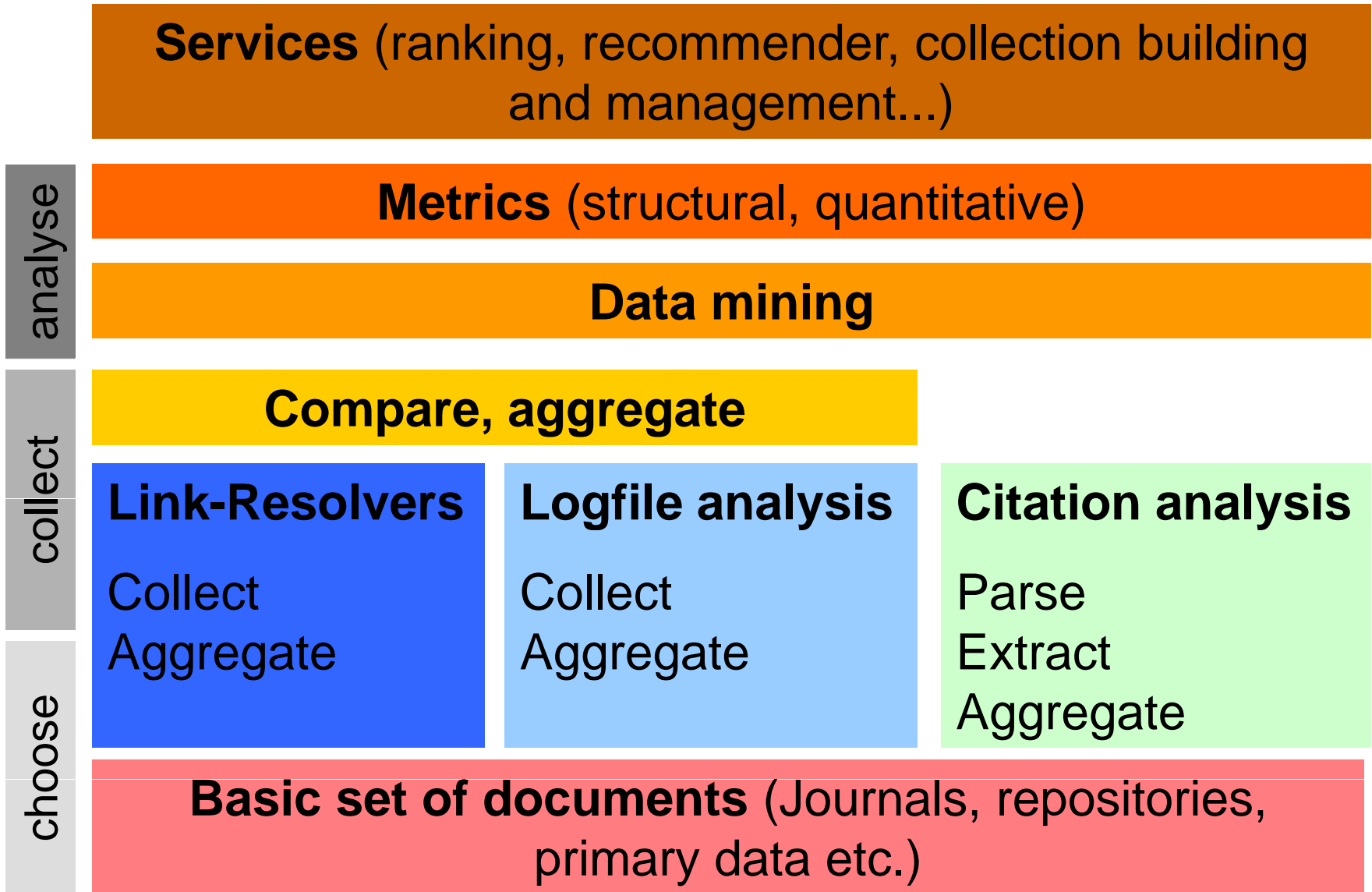


From: Bollen, Johan and Van de Sompel, Herbert (2005) A framework for assessing the impact of units of scholarly communication based on OAI-PMH harvesting of usage information. Delivered at OAI4, Geneva



# Measuring impact: the elements (schematic)

DEUTSCHE INITIATIVE FÜR NETZWERKINFORMATION E.V.





# Outline

1. Scope / definitions
2. Issues
3. Ongoing work / projects



## What are the basic items we want to count?

- Granularity (journals, articles, etc.)
- Beyond “publications” (primary data, learning objects, etc.)
- Identifying items (=referent deduplication)
  - Metadata based heuristics, persistent identifiers
- Identifying users (=agent deduplication)
  - User and session identification, pseudonymization



# Indicators and methods

- Log data processing
  - Identifying / filtering robots and crawlers
  - Proxies, caching
  - “click spans”
  - Grouping, isolating and aggregating useful usage patterns
  - Comparison and validation to citation data
- Aggregation and scalability
  - Technical issues
    - Different architectural frameworks: linking server-based, other, scalability, pseudonymization
  - Social/Policy issues
    - Networks of trust, sharing usage data
  - How is data collated with external sources (publisher data ...)



## Confidence in data and stats?

- Data validity
  - Usage definition, recording and representation, quality benchmarks, falsification issues
- Fraud and data manipulation
  - Standardized and transparent collection and aggregation of usage data
  - Auditing standards and procedures



## Legal issues, metrics and services

- Privacy and other legal issues
  - User and session identification, pseudonymization
  - Legal implications of log storage and aggregation
- Metrics and services
  - Investigating and tailoring metrics
  - Interfaces with existing bibliometric products
  - Definition of end-user services



# Background Issues

- What is the context for talking about impact?
- Why is promoting enhanced and alternative metrics important?
  - as a licensing issue: consortia and publishers want to investigate value for money issues
  - as a research policy issue:
    - scholarly communication changes rapidly
    - impact in new OA environment
    - trend analysis (“in-progress” communication)
    - assessment and evaluation of research



# Outline

1. Scope / Definitions
2. Issues
3. Ongoing work / projects



# Knowledge Exchange Workshop on Institutional Repositories

## Practical definition of 'usage'

- Raw format for exchanging usage events

## Items to be counted

- Lobby COUNTER to add article level stats

- Other academic "output types"

## Standard reports

- Agree on a small set of standard useful statistical reports that repositories should produce

## Policies for stats

- Compliance with local laws on e.g. privacy

- Enhance SHERPA policy tool

## Collection and aggregation

- Normalisation

- Specify issues of aggregation and deduplication for later study

## Collation with external sources

- Aggregating COUNTER stats at consortium level

- Investigate SUSHI interoperability with repositories and OAI-PMH + OpenURLContextObjects



## Ongoing work

- LANL
  - bX (with CalState, ExLibris)
  - MESUR
  - ...
- UK
  - University of Southampton (IRS, EPStats ...)
  - University College London
  - ...
- Germany (DINI / DFG)
  - Göttingen State and University Library
  - Stuttgart University Library
  - Computer and Media Service Humboldt University Berlin
  - Saarbrücken State and University Library



## Germany

- Workshop on Enhanced and Alternative Metrics of Publication Impact, 20–21 February 2006, Humboldt University Berlin
- Cluster of proposals to the DFG
  - Network of certified open access repositories 2y ✓
  - Usage statistics demonstrator ✓
  - Distributed open access reference citation service demonstrator ✓

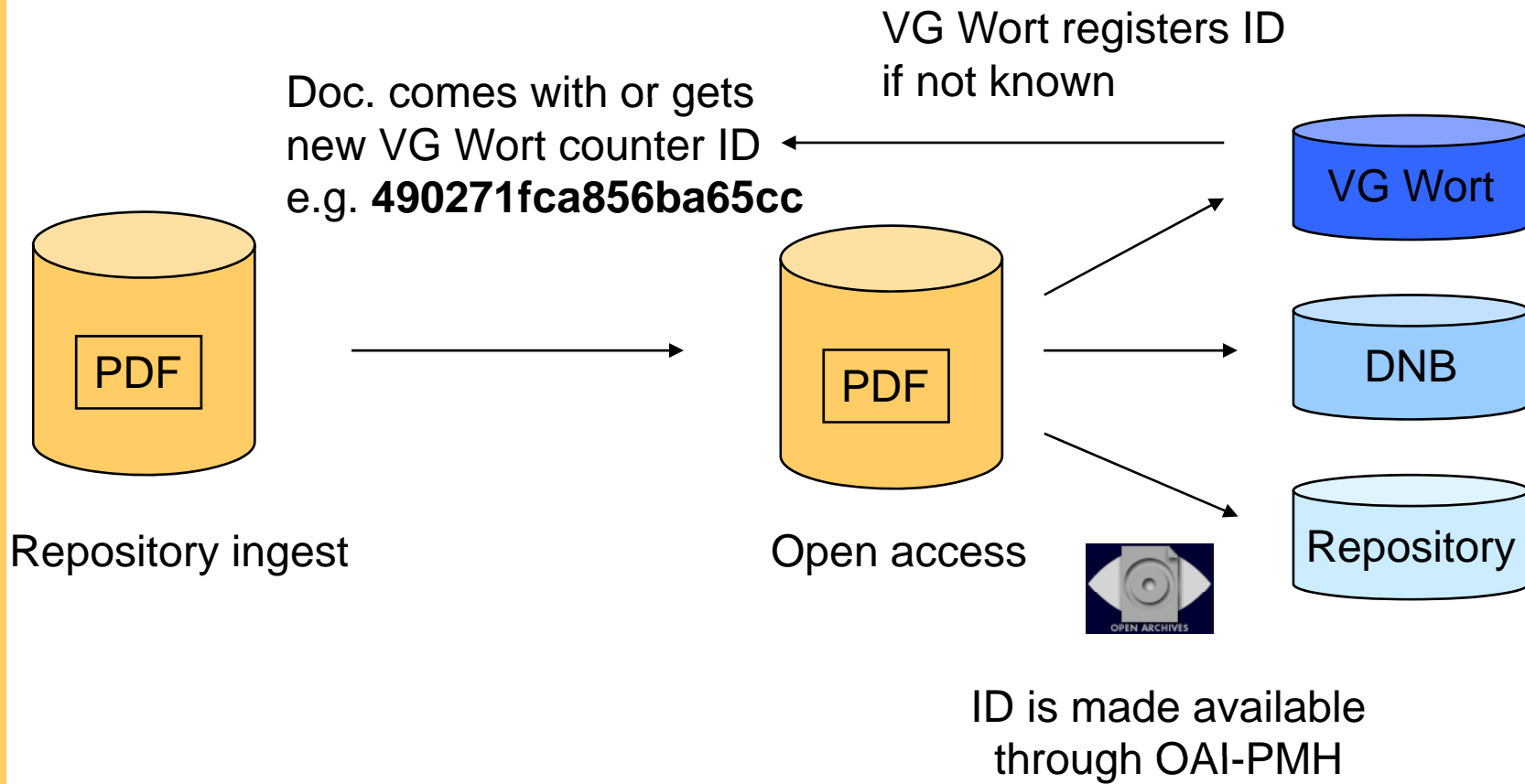


## Germany II

- German collecting society for copyright charges (VG Wort) has started a project on statistics (METIS)
- IFABC (International Federation of Audit Bureaux of Circulation) standards for robot detection and click spans (30 min)

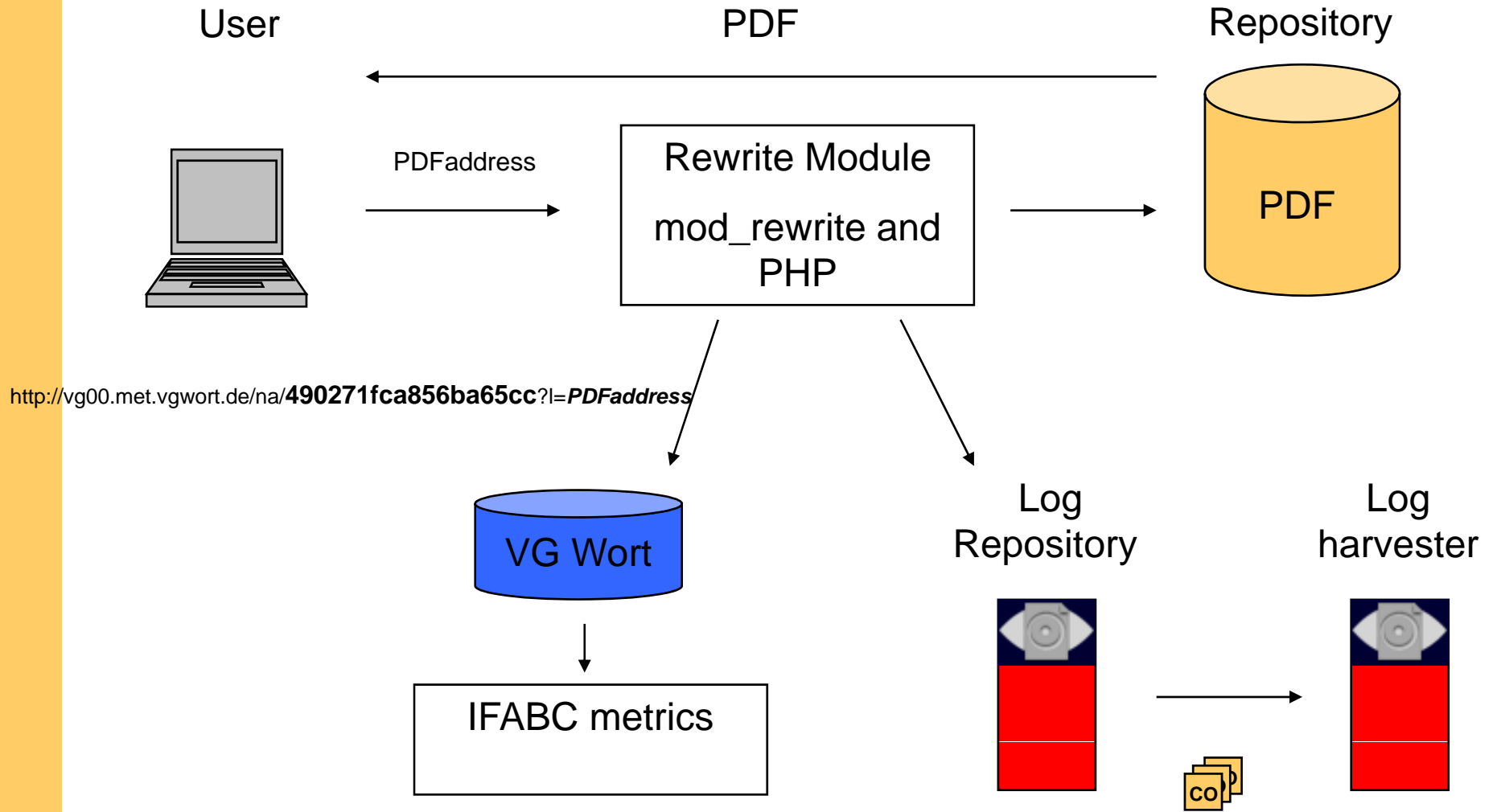


# Distributing VG Wort IDs



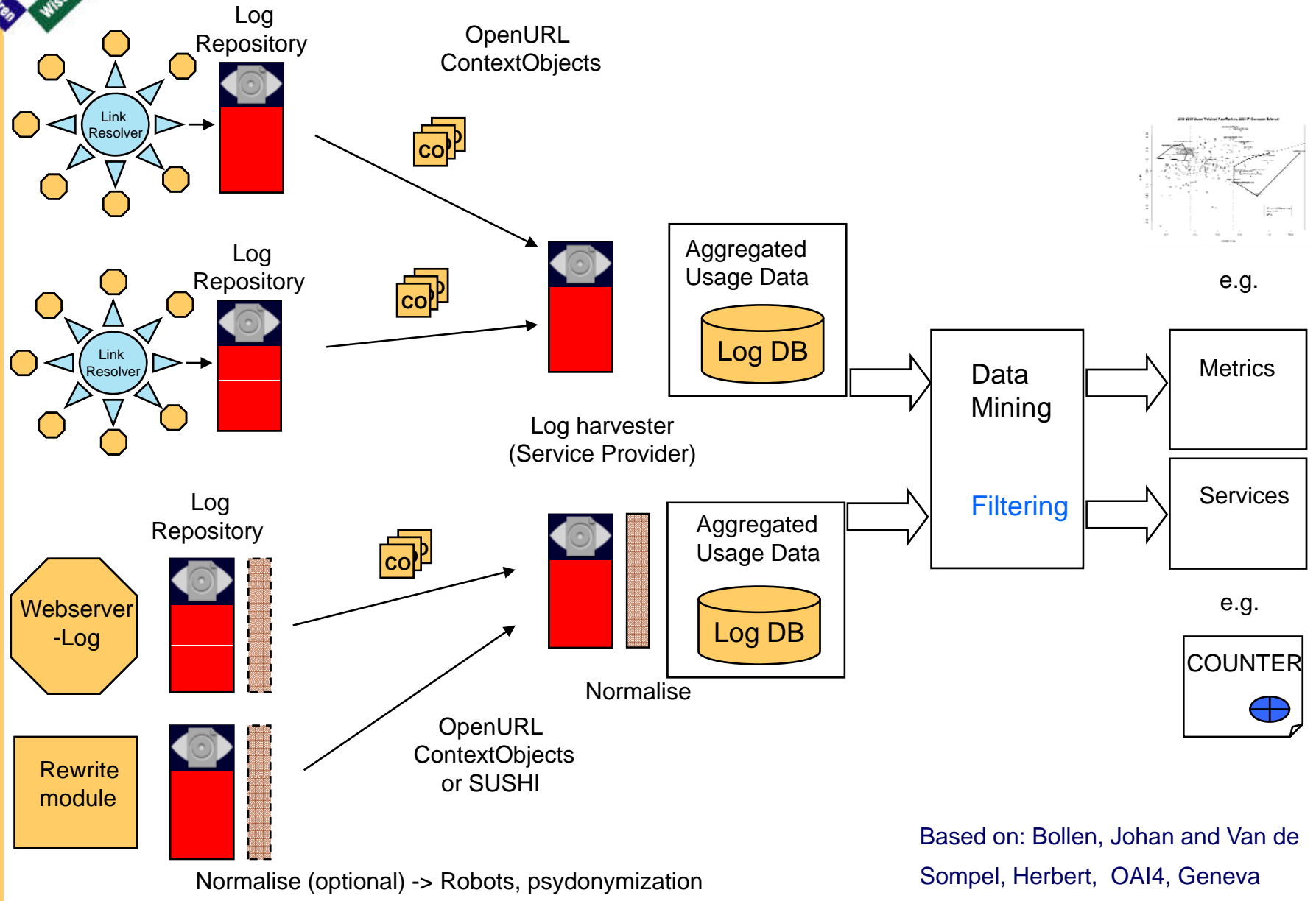


# Collecting data





# Measuring impact: the technical elements for usage data



Based on: Bollen, Johan and Van de Sompel, Herbert, OAI4, Geneva



## Conclusion

- Infrastructure for collecting and aggregating usage data is conceptually available, has to be deployed and implemented in practice on a large scale
- Investigating metrics for different needs and purposes
- Technical and social issues (the later posing the bigger challenge)
  - Standardization
  - Modularity
  - Co-Operation
  - Openness
  - Transparency