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"?
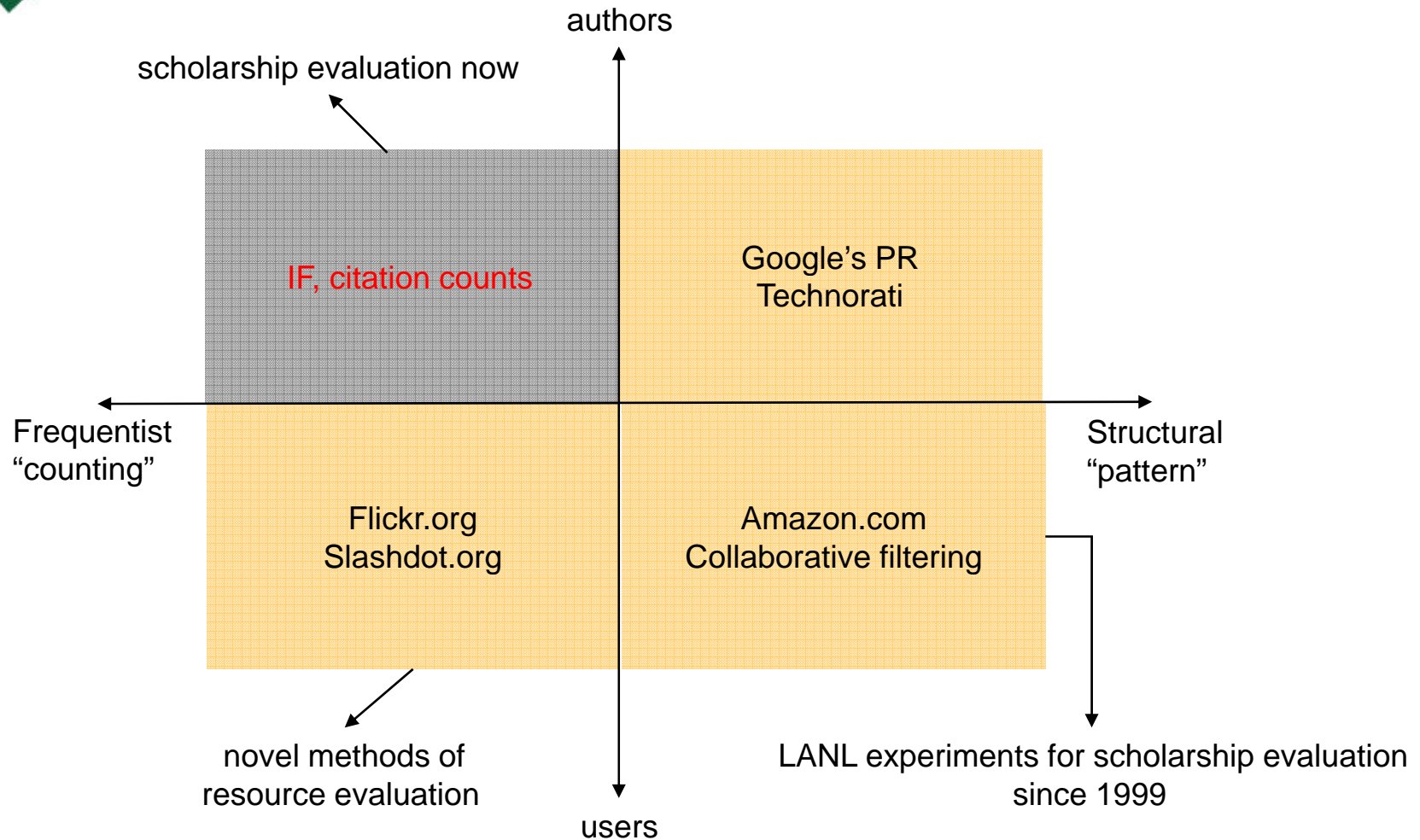
Measuring II

- 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

authors

scholarship evaluation now

IF, citation counts

Google's PR
Technorati

Frequentist
"counting"

Structural
"pattern"

Flickr.org
Slashdot.org

Amazon.com
Collaborative filtering

novel methods of
resource evaluation

users

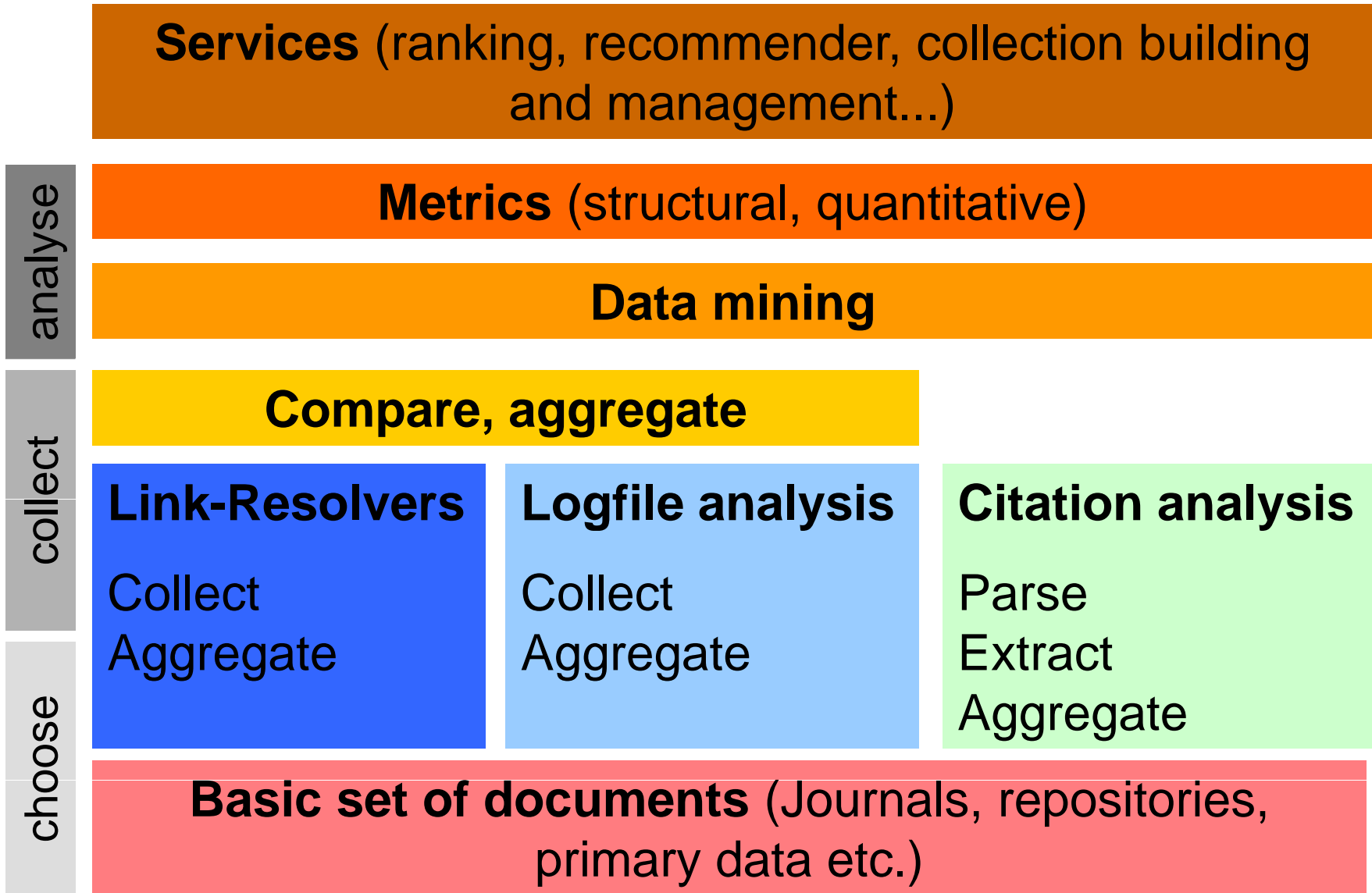LANL experiments for scholarship evaluation
since 1999

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)

Services (ranking, recommender, collection building and management...)

analyse

Metrics (structural, quantitative)

Data mining

collect

Compare, aggregate

choose

Link-Resolvers	**Logfile analysis**	**Citation analysis**
Collect	Collect	Parse
Aggregate	Aggregate	Extract
		Aggregate

Basic set of documents (Journals, repositories, primary data etc.)

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

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 ✓
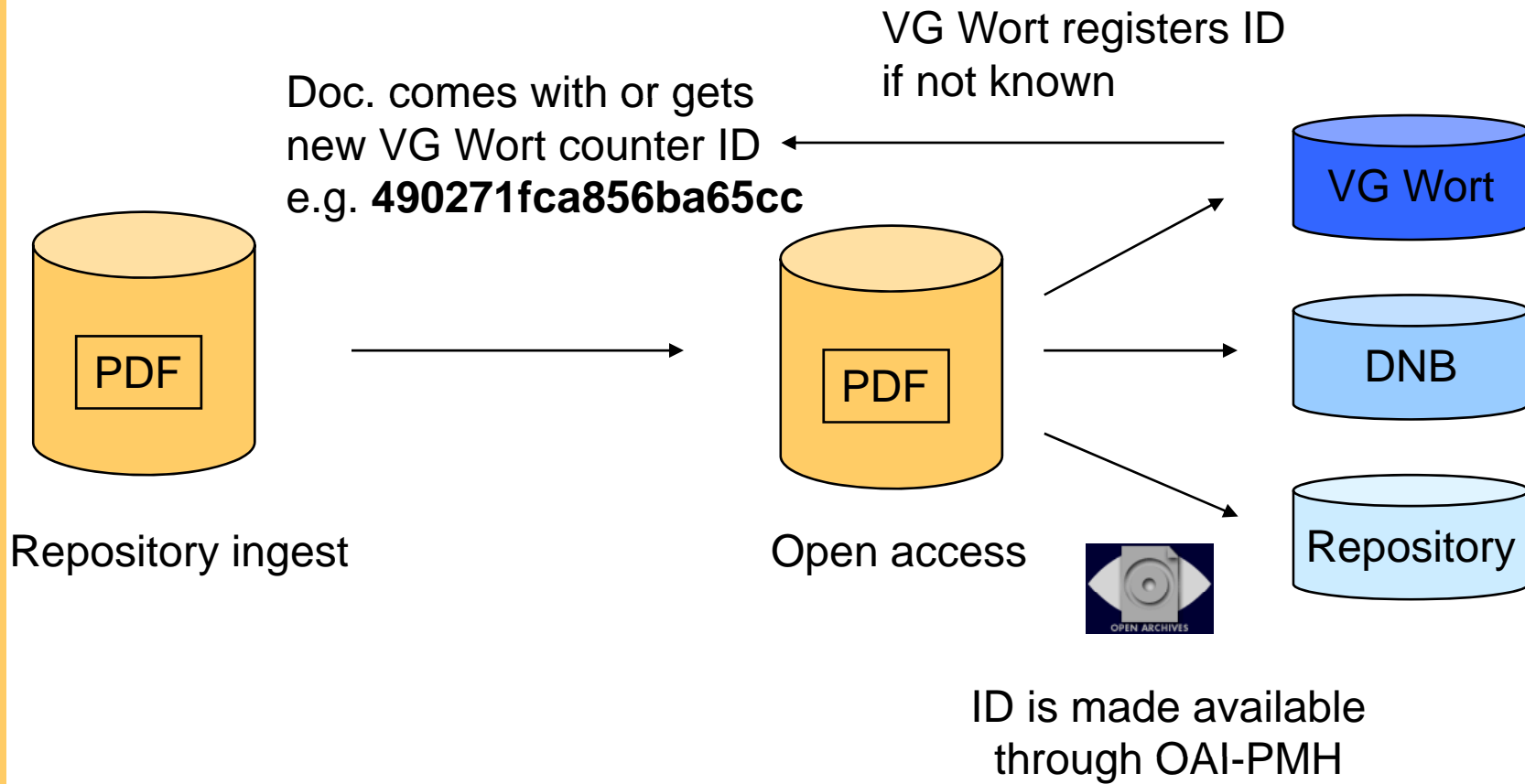 - Distributed open access reference citation service demonstrator ✓

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)
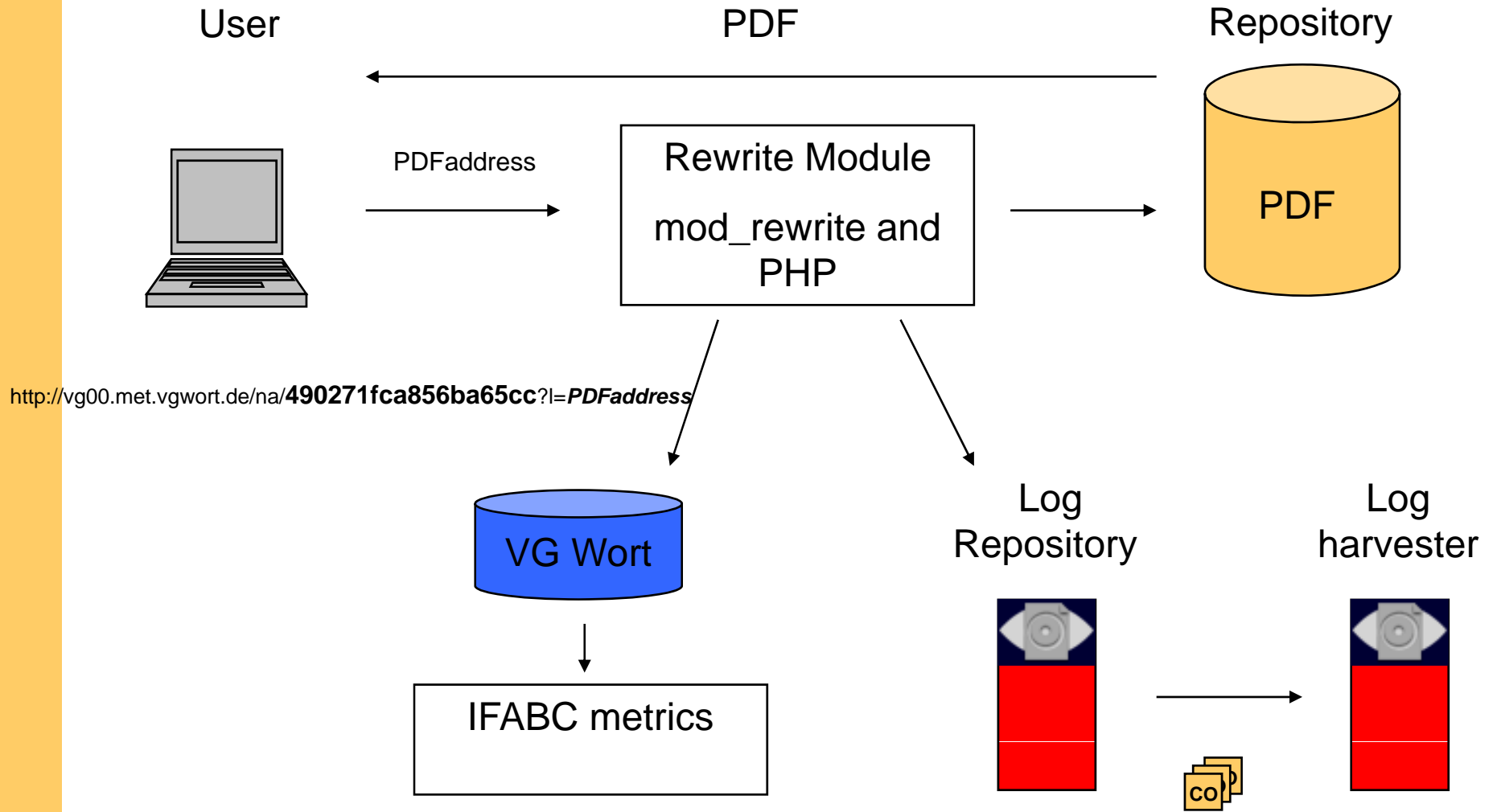
Distributing VG Wort IDs

Doc. comes with or gets
new VG Wort counter ID
e.g. **490271fca856ba65cc**

VG Wort registers ID
if not known

PDF

PDF

VG Wort

DNB

Repository

Repository ingest

Open access

ID is made available
through OAI-PMH

Collecting data

User

PDF

Repository

PDFaddress

Rewrite Module

mod_rewrite and PHP

PDF

http://vg00.met.vgwort.de/na/**490271fca856ba65cc**?l=*PDFaddress*

VG Wort

IFABC metrics

Log Repository

Log harvester

co

Measuring impact:
the technical elements for usage data

DEUTSCHE INITIATIVE FÜR NETZWERKINFORMATION E.V.

Log Repository

OpenURL ContextObjects

Link Resolver

co

Log Repository

Link Resolver

co

Aggregated Usage Data

Log DB

Log harvester (Service Provider)

Data Mining

Filtering

Metrics

e.g.

Log Repository

Webserver -Log

co

Normalise

Aggregated Usage Data

Log DB

Services

e.g.

Rewrite module

OpenURL ContextObjects or SUSHI

COUNTER

Normalise (optional) -> Robots, psydonymization

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