

# Repositories and Scholarly Communication

Herbert Van de Sompel

Digital Library Research & Prototyping Team  
Research Library  
Los Alamos National Laboratory, USA



Herbert Van de Sompel  
eSci-DR Workshop, September 3<sup>rd</sup>, Lisbon, Portugal



# Revolutionized Scholarship

Abstract for an August 29th 2007 Seminar at the Santa Fe Institute

SSWAP: Simple Semantic Web Architecture and Protocol

Damian Gessler (National Center for Genome Resources)

SSWAP (Simple Semantic Web Architecture and Protocol; pronounced "swap") is an architecture, protocol, and running platform for semantically integrating disparate data and services (see <http://sswap.info>, as in "swap info"). SSWAP is currently driving bioinformatic integration in the plant community under the Virtual Plant Information Network (<http://vpin.ncgr.org>), an NSF-funded semantic web services project. As an architecture, SSWAP establishes how data, service, and ontology providers, as well as clients and discovery servers can interact to allow for the description, querying, discovery, invocation, and response of semantic web services.




The 2007 Microsoft eScience Workshop at RENCI

https://www.mses07.net/main.aspx

rencl esience

Connotea



Home Registration Agenda Logistics

## The 2007 Microsoft eScience Workshop at RENCI

---

**Deadline for abstracts extended to Monday August 20, 12 noon PST**

Please submit abstracts online at <https://cmt.research.microsoft.com/escience07/>  
Authors will be notified of acceptance by **Monday September 3rd**.  
(Only abstracts from registered attendees will be accepted; please register using the link at the bottom of this page)

---

**October 21-23 2007**  
**The Friday Center for Continuing Education**  
UNC - Chapel Hill  
100 Friday Center Drive  
Chapel Hill, NC 27599-1020

**It is no longer possible to do science without computing.**

The use of computers creates many challenges as it expands the realm of the possible in scientific research and many of these challenges are common to researchers in different areas. The insights gained in one area may catalyze change and

<http://www.mses2007.net/>



Herbert Van de Sompel  
eSci-DR Workshop, September 3<sup>rd</sup>, Lisbon, Portugal



# Revolutionized Scholarship

Multiple accelerating trends are converging and crossing thresholds in ways that show extraordinary promise ... in how we create, disseminate, and preserve scientific and engineering knowledge.

Dramatic new capabilities:

- Raw computing power
- Storage capacity
- Algorithms
- Network throughput
- Measurement techniques
- Data-mining techniques

[http://www.communitytechnology.org/nsf\\_ci\\_report/](http://www.communitytechnology.org/nsf_ci_report/)



# And Scholarly Communication?

The established scholarly communication system has not kept pace with these revolutionary changes in research practice.

- The current electronic scholarly communication system is to an extent a scanned copy of the paper-based system.
- The networked environment begs for another scholarly communication system to emerge.
- Let's build the system scholars deserve!



Rethinking Scholarly Communication: Building the System that Scholars Deserve - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.dlib.org/dlib/september04/vandesompel04-vandesompel.html

Getting Started Latest Headlines

Search | Back Issues | Author Index | Title Index | Contents

---

**OPINION**

---

**D-Lib Magazine**  
September 2004

Volume 10 Number 9  
ISSN 1082-9873

**Rethinking Scholarly Communication**  
**Building the System that Scholars Deserve**

[Herbert Van de Sompel](#)  
Los Alamos National Laboratory, Research Library  
<herbertv@lanl.gov>

[Sandr Payette](#)  
Cornell University, Computing and Information Science  
<payette@ci.cornell.edu>

[John Enckeson](#)  
Hewlett-Packard Laboratories, Digital Media Systems Lab  
<john.enckeson@hp.com>

[Carl Lagoze](#)  
Cornell University, Computing and Information Science  
<lagoze@ccornell.edu>

Done

Start | 2 files... | Calendar... | Keynote | 2 files... | Firefox... | Yahoo... | 2:47 PM

<http://dx.doi.org/10.1045/september2004-vandesompel>



Herbert Van de Sompel  
eSci-DR Workshop, September 3<sup>rd</sup>, Lisbon, Portugal



# Revolutionizing Scholarly Communication

- Repositories
- Value Chains Starting in Repositories
- Interoperability
- Rights
- Machine Readability
- Metrics



# Revolutionizing Scholarly Communication

- Repositories
  - Value Chains Starting in Repositories
  - Interoperability
  - Rights
  - Machine Readability
  - Metrics



# Repositories

- Preprint repositories,
- Publisher repositories,
- Postprint repositories,
- Dataset repositories,
- Software repositories,
- Cultural heritage collections,
- Learning & Teaching object repositories,
- Digitized book repositories,
- ....

Can be institution-based, discipline-based, corporate, ...



# Repositories

- Quite difficult to define what repositories exactly are.
- But:
  - Networked service
  - Ingests digital scholarly materials
  - Hosts those materials
  - Manages those materials
  - Have some notion of a long-term horizon (although we don't know how to make this a reality)
  - Makes those materials accessible
    - What's in a word!



# Revolutionizing Scholarly Communication

- Repositories
- Value Chains Starting in Repositories
- Interoperability
- Rights
- Machine Readability
- Metrics



# Value Chains Starting in Repositories

- We must leverage the value of the materials that become available in those distributed Repositories.
- Think about these Repositories as active nodes in a global environment, not as passive local nodes
  - These Repositories are **not** about creating **services for local users** (only)
  - These Repositories are **not** about creating **a service** (user interface) **for all users**



# Value Chains Starting in Repositories

- Repositories as the foundation for the future scholarly communication system.
- Materials from Repositories must be re-usable in different contexts.
- Life for those materials starts there, not ends there.



# Revolutionizing Scholarly Communication

- Repositories
- Value Chains Starting in Repositories
- Interoperability
- Rights
- Machine Readability
- Metrics



# Interoperability

- You like what you can do in Flickr, YouTube, etc?
  - Social stuff like comment, rate, reference, blog about, embed, ...
  - Yeah, like ... err ... communication?
- We need to be able to do exactly those things with scholarly materials.
- And, just like discovery at the level of a single repository is meaningless, so is the implementation of social tools stove-piped upon a single repository.
- We need to be able to do those things across repositories and across networked applications.
- That requires interoperability.
  - This has many facets!



# Interoperability : OAI Object Re-Use & Exchange

- Problem: New scholarly communication materials are “compound”:
  - Multiple datastreams
  - Multiple media types
  - Multiple semantic types
  - Multiple network locations
- How to reference them?
- How to reference their components in a manner that makes clear they are part of a compound object?
- How to re-use them (ship them around) without actually having to ship the real content around?
- And how to do so in a manner that fully leverages the Web architecture?



# Interoperability : OAI Object Re-Use & Exchange

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

# Interoperability : OAI Object Re-Use & Exchange

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

ixi



## Open Archives Initiative Object Reuse and Exchange

Home

Projects

Specifications

Community

About OAI

Open Archives Initiative -> ORE

### Exchanging Information about Digital Objects

ORE will develop specifications that allow distributed repositories to exchange information about their constituent digital objects. These specifications will include approaches for representing digital objects and repository services that facilitate access and ingest of these representations. The specifications will enable a new generation of cross-repository services that leverage the intrinsic value of digital objects beyond the borders of hosting repositories.

### The OAI-ORE Community

- [Executive Committee](#)
- [Advisory Committee](#)
- [Technical Committee](#)
- [Liaison Group](#)

### Contact us

- [ore@openarchives.org](mailto:ore@openarchives.org)

### Selected OAI-ORE Resources

- [Report of the May 2007 ORE-TC Meeting](#) - The focus of the meeting was building consensus on named graphs as a underlying model for expressing compound object information.
- [Compound Information Objects white paper](#)  
A white paper describing the web-centric OAI-ORE perspective on compound information objects is now available. This document is a work in progress and was used as a discussion document in preparation of the May 2007 meeting of the OAI-ORE Technical Committee.
- [Proposal for funding to the Mellon Foundation](#)  
Details the plan for work developing OAI-ORE specifications over the two-year period beginning October 2006
- [Augmenting interoperability across scholarly repositories](#)  
An April 20-21 2006 meeting sponsored by Microsoft, CNI, DLF, and JISC that laid the foundation for OAI-ORE
- [Rethinking Scholarly Communication: Building the System that Scholars Deserve](#)  
An opinion article in D-lib that describes a model of a scholarly communication system that interoperable repositories could provide
- [All OAI-ORE Resources ...](#)

<http://www.openarchives.org/ore/>



Herbert Van de Sompel  
eSci-DR Workshop, September 3<sup>rd</sup>, Lisbon, Portugal



# CTWatch QUARTERLY

CTWatch Quarterly Current Issue | CTWatch Quarterly Archives | About CTWatch Quarterly

AUGUST 2007

VOLUME 3 NUMBER 3

THE COMING REVOLUTION IN SCHOLARLY COMMUNICATIONS & CYBERINFRASTRUCTURE

Printable Format

## Interoperability for the Discovery, Use, and Re-Use of Units of Scholarly Communication

**Herbert Van de Sompel**, Los Alamos National Laboratory  
**Carl Lagoze**, Cornell University

1

### 1. Introduction

Improvements in computing and network technologies, digital data capture, and data mining techniques are enabling research methods that are highly collaborative, network-based, and data-intensive. These methods challenge existing scholarly communication mechanisms, which are largely based on physical (paper, ink, and voice) rather than digital technologies.

One major challenge to the existing system is the change in the nature of the *unit* of scholarly communication. In the established scholarly communication system, the dominant communication units are journals and their contained

<http://www.ctwatch.org/quarterly/articles/2007/08/interoperability-for-the-discovery-use-and-re-use-of-units-of-scholarly-communication>



Herbert Van de Sompel  
eSci-DR Workshop, September 3<sup>rd</sup>, Lisbon, Portugal



# Revolutionizing Scholarly Communication

- Repositories
- Value Chains Starting in Repositories
- Interoperability
- Rights
- Machine Readability
- Metrics



# Rights

- Machines are the next generation readers.
- They must be able to understand what can be done with the materials they discover.
- So many wonderful things can happen when we tell machines what they are allowed to do.
  - process information to extract knowledge, attach knowledge, mine, evolve, build upon
- Need for an environment in which scholarly assets behave in a manner that matches the “gift exchange” spirit of scholarship.



Science Commons

http://sciencecommons.org/

Connotea

science commons

THE SCIENCE COMMONS  
PROJECTS  
RESOURCES  
PARTNERS  
CONTACT/BLOG

**Accelerating the Scientific Research Cycle**

Science Commons serves the advancement of science by removing unnecessary legal and technical barriers to scientific collaboration and innovation.

Built on the promise of Open Access to scholarly literature and data, Science Commons identifies and eases key barriers to the movement of information, tools and data through the scientific research cycle.

Our long term vision is to provide more than just useful contracts.

**SC Blog**

[SciVee round-up](#)

So much has been written about SciVee, the new Web site dubbed "the YouTube for science research papers", that we've decided to take on the round-up post instead of restating the good words of our peers. For those who may not know, the site is a project of the National Science Foundation, the San Diego I. 1

<http://www.sciencecommons.org>



Herbert Van de Sompel  
eSci-DR Workshop, September 3<sup>rd</sup>, Lisbon, Portugal





Science Commons » Neurocommons Technical Overview

http://sciencecommons.org/projects/data/nc\_technical\_ow

mons neurocommons

Connotea



---

## Neurocommons Technical Overview

---

(The following technical pages are mirrored at [sw.neurocommons.org](http://sw.neurocommons.org))

The success of a scientific investigation is determined in part by its ability to locate and make effective use of relevant prior work. Automated literature search is a basic tool used by all scientists, but the computer and the Internet have potential for search and integration far beyond what can be done with keyword-based search. However, a prerequisite for automated exploitation of scientific information is that it be in a consistent format that can be processed meaningfully and accurately by software. We need links among literature, data records, real-world entities, and abstract concepts, with formal definitions of each link's endpoints and type. Applications need to use common identifiers for endpoints so that mentions of shared entities can be matched. This discipline of links, definitions, and identification is exactly what the framework of the semantic web provides.

In 2007 Science Commons intends to roll out artifacts and demonstrations that show the construction of a semantic web for science – in particular, for neuroscience. Our efforts toward such a "Neurocommons" are in three areas:

1. Data integration
2. Text mining
3. Analytic tools

[Neurocommons Data Integration Pilot](#)

[Neurocommons Text Mining Pilot](#)

[The Science Commons Projects](#)

[Resources](#)

[Partners](#)

[Contact / Blog](#)

[SC Blog](#)

<http://sciencecommons.org/projects/data/>



Herbert Van de Sompel  
eSci-DR Workshop, September 3<sup>rd</sup>, Lisbon, Portugal



# Revolutionizing Scholarly Communication

- Repositories
- Value Chains Starting in Repositories
- Interoperability
- Rights
- Machine Readability
- Metrics



# Machine Readability

- Need machine readable primary information
- A remarkable amount of information is lost when it is being communicated in a “paper”:
- But I don’t need to explain this. Peter Murray-Rust can do that so much better.



cml.sourceforge.net

http://cml.sourceforge.net/ cal markup language

Connotea

Updated: April 15, 2007. XML

## cml.sourceforge.net - OpenSource Site for CML

- Home
- Schemas
- Mailing list
- Bibliography
- Downloads
- Other Software
- This project
- Related projects
- What's New
- Historical

Search

### What's New

- 15.04.07: Updated bibliography.
- 13.01.06: A [CML Wiki](#) to discuss and disseminate news and information is now available
- 7.02.05: CMLReact. A publication-ready new version of this [schema is now available](#).
- 6.09.04: Chemical Markup Language: [a two-day workshop at the University of Cambridge](#).
- 15.03.04: CMLRSS is a new "newsfeed" protocol using CML as an extension to the RDF-based RSS 1.0 protocol. Distribution kits are available [here](#) or [here](#) and the method is published as: P. Murray-Rust, H. S. Rzepa, M. J. Williamson and E. L. Willighagen, "Chemical Markup, XML and the Worldwide Web. Part 5. Applications of Chemical Metadata in RSS Aggregators", *J. Chem. Inf. Comp. Sci.*, 2004, **44**, 462-469.
- 1.03.04: [cdx2cml \(ChemDraw\) and Word2cml Legacy converters \(Java source code\) available](#).
- 18.01.04: A Complete [Element and Attribute](#) listing for the CML Family (generated from the schemas, 724 entries).
- 1.12.03. The [Complete CML Schema](#) has been added to the Wiki pages
- 3.11.03. [A Wiki page has been set up](#) for collaborative information regarding the CML language family.
- 1.11.03. Editorial in [Nature Structural Biology](#) - highlighting CML.
- 21.09.03. The CML1 and CML2 processors in [openBabel](#) have been updated to correct various bugs in the CML parsing. OpenBabel 1-100-1 now incorporates these changes which we have compiled into binaries (executables) for Openbabel as follows
  1. [MacOS X \(10.2 and 10.3\)](#)
  2. [SGI \(Irix 6.5\)](#)
  3. [Linux \(Redhat 8\)](#)
  4. [Windows \(install both babel.exe and cygwin1.dll in same directory\)](#).
  5. Examples of CML1 and CML2 are [available here](#), along with the Molfiles they were generated from.Openbabel is invoked only from the [command line](#) (there is no GUI version). A copy of the cml.cpp module is [available here](#) if you wish to compile your own version of openbabel. Please download the versions as of 21.09.03 rather than earlier versions.
- 14.07.03. The CMLCore schema is updated to version 2.1.1 correcting various validation bugs.
- 8.07.03. A collection of Scientific Markup languages is described at [Robin Cover's Materials Page](#).
- 24.05.03. Part 4 of the series on CML is published as: P. Murray-Rust and H. S. Rzepa, "Chemical Markup, XML and the Worldwide Web. CML Schema", *J. Chem. Inf. comp. Sci.*, 2003, **757** - 772.
- 22.11.02. A discussion list for CML, CMLComp (Computational Chemistry Markup Language), STMML and other relatives of this family is available at <https://lists.sourceforge.net/lists/listinfo/cml-discuss>
- 19.11.02. CML will be presented at two invited talks submitted to the [Division of Chemical Information](#) for the 225th ACS National Meeting, New Orleans, LA, March 23-27, 2003 in New Orleans (from

http://cml.sourceforge.net



Herbert Van de Sompel  
eSci-DR Workshop, September 3<sup>rd</sup>, Lisbon, Portugal



Untitled Document

http://neuroscientific.net/sisc/sisc\_introduction.htm

Connotea

# Semantically Interlinked Scientific Communities

Matthias Samwald, June 2007, draft

The *Semantically Interlinked Scientific Communities* project (from here on called *SISC*) is an ambitious project that uses Semantic Web standards (RDF, OWL, RDFa) to drastically improve how scientific data and knowledge is represented and communicated on the web. It builds on established ontologies and metadata standards and adapts them for the use in scientific practice.

Features:

- \* Structured, semantic representation of scientific discourse. With SISC it is possible to make statements about agreement, disagreement and other relations between scientific documents, research findings, database entries, and even between different portions of text.
- \* The barriers and distinctions between publications and databases are removed. RDF/OWL is embedded directly into the text -- not as a mere annotation of the text, but as a direct representation of biological reality.
- \* Fine-grained control of authorship information and copyright status of documents and database entries.
- \* Direct integration of most important biomedical ontologies (e.g. from the Open Biomedical Ontologies collection).
- \* SISC transforms popular web communication platforms like weblogs and bulletin boards into perfect tools for scientific discourse. The flow of information between different e-mails, blog entries and forum postings is captured through explicit semantic structures and connects seamlessly to scientific publications and database entries. All of this can be queried as a coherent Semantic Web.
- \* Bibliographic, personal and organizational information is expressed through Semantic Web standards that ease personal information management.

Some basic (and very incomplete) example pages that demonstrate the use of SISC in research on Alzheimer's disease:

[Publication 1](#)  
 Publication 2 (motivated by publication 1)  
 Publication 3 (contains statements that conflict with publication 1)  
[Example of a scientific weblog](#) containing RDF metadata generated with the SIOC plugin for Wordpress. The resulting RDF looks like [this](#) (automatically produced with the SIOC browser).

## The building blocks of SISC

SIOC	“SIOC provides methods for interconnecting discussion methods such as blogs, forums and mailing lists to each other. It consists of the SIOC ontology, an open-standard machine readable format for
Semantically Interlinked Open	expressing the information contained both explicitly and implicitly in internet discussion methods, of

[http://neuroscientific.net/sisc/sisc\\_introduction.htm](http://neuroscientific.net/sisc/sisc_introduction.htm)



# Revolutionizing Scholarly Communication

- Repositories
- Value Chains Starting in Repositories
- Interoperability
- Rights
- Machine Readability
- Metrics



# Metrics

- Evaluation of scholarly status matters:
  - 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
- But a single metric rules



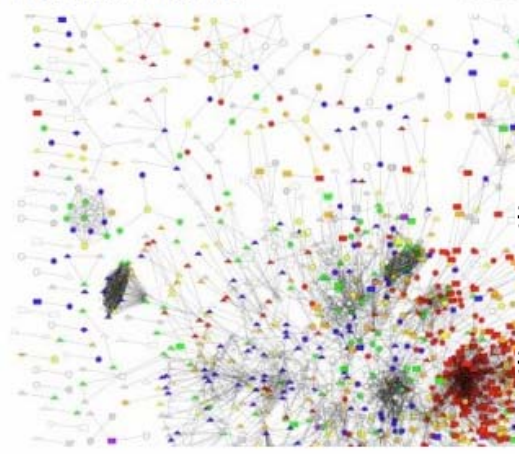
# Metrics

- We urgently need new metrics that:
  - Are able to take into account the breadth new scholarly materials, not just journal articles
  - Are based on network recordable/extractable data





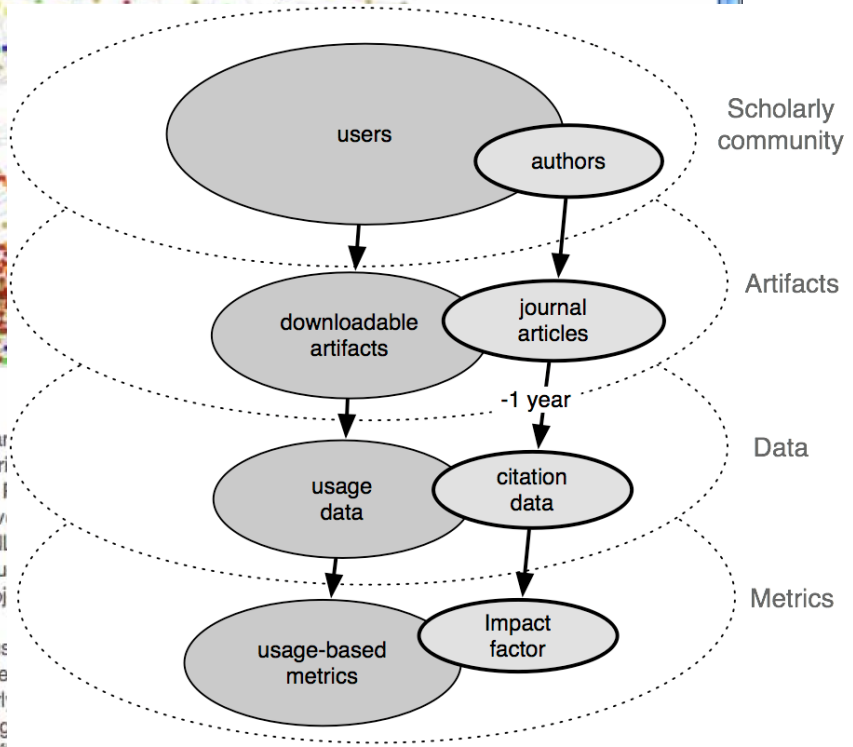
MEtrics from Scholarly Usage of Resources  
 Los Alamos National Laboratory



**Overview:**

The Andrew W. Mellon Foundation has awarded a grant support of a two-year project that will investigate metrics from scholarly information. The Digital Library Research & I carry out the project. Johan Bollen is the Principal Investigator, architectural consultant, and Aric Hagberg of the LANL as modeling consultant. Marko A. Rodriguez, PhD student at LANL Graduate Research Assistant, supports the project.

The project's major objective is enriching the toolkit used for scholarly communication items, and hence of scholars, with metrics. It starts with the creation of a semantic model of scholarly information, a semantic store that relates a range of scholarly bibliographic



<http://www.mesur.org>

