Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive, 14 - 38123 POVO, Trento - Italy
http://disi.unitn.it

STATE OF THE ART IN SCIENTIFIC KNOWLEDGE CREATION, DISSEMINATION, EVALUATION AND MAINTENANCE

Maintainer/Editor-in-chief	Aliaksandr Birukou
Core authors	Marcos Baez, Aliaksandr Birukou, Ronald Chenu, Matus Medo, Nardine Osman, Diego Ponte, Jordi Sabater-mir, Luc Schneider, Matteo Turrini, Giuseppe Veltri, Joseph Rushton Wakeling, Hao Xu
Maintainers/Editors	Aliaksandr Birukou, Ronald Chenu, Jim Law, Nardine Osman, Diego Ponte, Azzurra Ragone, Luc Schneider, Matteo Turrini
LiquidPub research group leaders	Fabio Casati, Roberto Casati, Ralf Gerstner, Fausto Giunchiglia, Maurizio Marchese, Gloria Origgi, Alessandro Rossi, Carles Sierra, Yi-Cheng Zhang
LiquidPub project leader	Fabio Casati

December 2009

Technical Report # DISI-09-067

D1.1
State of the Art

Lead partner: UNITN; Contributing partners: IIIA-CSIC, CNRS, UNIFR, SPRINGER

Grant agreement no.	LiquidPub/2009/D1.1/v2.0
Project acronym	EU FET OPEN: FP7-ICT-2007-C
Version	v2.0
Date	December 18, 2009
State	Solid
Distribution	Public

Disclaimer

This document is part of a research project funded by the European Community as project number 213360 acronym LiquidPublication under THEME 3: FP7-ICT-2007-C FET OPEN. The full list of participants is available at `http://project.liquidpub.org/partners-and-contributors/liquidpub-teams`.

Abstract

This report presents an overview of the State of the Art in topics related to the on-going research in the Liquid Publications Project. The Liquid Publications Project (LiquidPub) aims to bring fundamental changes to the processes by which scientific knowledge is created, disseminated, evaluated and maintained. In order to accomplish this, many processes and areas will have to be modified. We group the areas involved in this change into four areas: creation and evolution of scientific knowledge, evaluation processes (primarily, peer-review processes and their evaluation), computational trust and reputation mechanisms, and business and process models. Due to the size and complexity of each of these four areas, we will only discuss topics that are directly related to our proposed research.

Keyword list: **knowledge artefacts**, **knowledge modelling**, **knowledge evolution**, **research assessment**, **peer review**, **trust**, **reputation**, **process models**, **copyright**, **licensing**, **open access**

Executive Summary

The ultimate goal of the Liquid Publications Project, LiquidPub, is to find new processes for the creation, evolution, evaluation, dissemination, and reuse of scientific knowledge in all fields. We take inspiration from the co-evolution of scientific knowledge artefacts and software engineering artefacts. We feel the paradigm changes brought about by the World Wide Web and Web 2.0 applications have not been applied to improve the scientific process nearly as much as they could be.

The processes involved in creating, evaluating, and disseminating scientific knowledge are as much social processes as they are technical processes. Therefore, LiquidPub has both social and technical concerns. In this State Of The Art document we will outline the major scientific fields and topic areas that we feel are critical to the research we propose to accomplish.

In the chapters of this report we examine four broad areas we see as related to our research on changing scientific knowledge processes.

- *Modelling evolutionary, collaborative, and multi-faceted knowledge artefacts.* The first section is a review of existing approaches to modelling and storing complex digital objects and their evolutions. Software versioning systems, wikis, and document repositories are examples of such systems.

- *Analysis of reviews and modelling of reviewers' behaviour in peer review.* This section reviews two lines of research: efforts that analyse review processes in scientific fields and approaches to modelling reviewer behaviour. Besides peer reviews and reviewers, the study also touches upon how reviews are done in other related fields where creative content is produced, such as evaluation of software artefacts, pictures, or movies. The task also identifies metrics for "good" reviews and review processes and how to measure them.

- *Computational trust, reputation models, and social network analysis.* Subsections include the latest computational trust and reputation models in the area and deeper analysis of those models that are relevant for our scenario, that is models that take into account social dimensions and models associated with web ratings such as PageRank. The section concludes with discussion of relevant work on social network analysis that can be useful in computational trust and reputation models.

- *Copyrights, licensing, and business models.* This section will review the current processes and practices in copyright and licensing of content in various areas, from software to art to scientific creation in general. It will also review approaches to copyright management. We also discuss state of the art in innovative publisher services for liquid publications, and how the publishing industry may be affected by the transition to a liquid publication model. Specifically, we analyse current intellectual property rights schemes and the industry value chain in the current publication model. We then identify innovative services that publishers can offer to the scientific community, to add value and hence generate business, for them.

Contents

Part 1

Introduction

The world of scientific publications has been largely oblivious to the advent of the Web and to advances in ICT. Even more surprisingly, this is the case for research in the ICT area. ICT researchers have been able to exploit the Web to improve the knowledge production process in almost all areas except their own. We are producing scientific knowledge (and publications in particular) essentially following the very same approach we followed before the Web. Scientific knowledge dissemination is still based on the traditional notion of "paper" publication and on peer review as quality assessment method. The current approach encourages authors to write many (possibly incremental) papers to get more "tokens of credit", generating often unnecessary dissemination overhead for themselves and for the community of reviewers. Furthermore, it does not encourage or support reuse and evolution of publications: whenever a (possibly small) progress is made on a certain subject, a new paper is written, reviewed, and published, often after several months. The situation is analogous if not worse for textbooks.

The Liquid Publications Project (LiquidPub) proposes a paradigm shift in the way scientific knowledge is created, disseminated, evaluated and maintained. This shift is enabled by the notion of *Liquid Publications*, which are evolutionary, collaborative, and composable scientific contributions. Many Liquid Publication concepts are based on parallels between scientific knowledge artefacts and software artefacts, and hence on lessons learned in (agile, collaborative, open source) software development, as well as on lessons learned from Web 2.0 in terms of collaborative evaluation of knowledge artefacts.

A critical enabling activity for this paradigm shift is a comprehensive understanding the state of the art in the topic areas related to the change. This document will outline current technology and thinking in the areas of creation, dissemination, evaluation, and reuse of scientific knowledge.

To introduce our areas of concern, consider the following academic scenario:

> A PhD student discusses with his advisor some new approaches in a specific field. They start to collaborate on a proposal for the scientific work and experiments. If successful, the work is submitted and accepted at a workshop and a website is created to publicly present the work. After comments from reviewers and attendees, additional research directions are investigated. Persons from the workshop with specific and related competences become involved. The article presented at the workshop is divided into two threads of work involving the original group and the other people and both

threads are submitted to relevant conferences.

One thread is accepted at a conference and the other is rejected. In both cases the reviewer's feedback is used to guide future work, particularly in the case of the rejected work. The rejected work is improved and submitted to another conference where it is accepted.

After the presentations, audience feedback, attendees' discussions, and further grounding work, an extended paper including parts from all previous publications (for example state-of-the-art, methodologies, experiments, results, new material, and analysis) is submitted with all the relevant authors and gets accepted in a suitable journal with high impact factor. Draft copies of the articles are made available to the public free of charge on a University website.

Meanwhile, other groups of researchers contact the initial group of authors and begin collaborating in several clusters along different lines of research stemming from the original initial idea. Eventually, the original researchers start a business to commercialize their ideas.

In this relatively common scenario we find most of the processes that the LiquidPub Project is concerned with: creation and evolution of scientific knowledge, collaboration, trust and reputation, peer-review, dissemination, intellectual property rights, and business models.

In the chapters of this report we examine the four broad areas we see as related to scientific knowledge processes.

- *Modelling evolutionary, collaborative, and multi-faceted knowledge artefacts.* The first section is a review of existing approaches to modelling and storing complex digital objects and their evolutions. Software versioning systems, wikis, and document repositories are examples of such systems.

- *Analysis of reviews and modelling of reviewers' behaviour in peer review.* This section reviews two lines of research: efforts that analyse review processes in scientific fields and approaches to modelling reviewer behaviour. Besides peer reviews and reviewers, the study also touches upon how reviews are done in other related fields where creative content is produced, such as evaluation of software artefacts, pictures, or movies. The task also identifies metrics for "good" reviews and review processes and how to measure them.

- *Computational trust, reputation models, and social network analysis.* Subsections include the latest computational trust and reputation models in the area and deeper analysis of those models that are relevant for our scenario, that is models that take into account social dimensions and models associated with web ratings such as PageRank. The section concludes with discussion of relevant work on social network analysis that can be useful in computational trust and reputation models.

- *Copyrights, licensing, and business models.* This section will review the current processes and practices in copyright and licensing of content in various areas, from software to art to scientific creation in general. It will also review approaches to copyright management. We also discuss state of the art in innovative publisher services for liquid publications, and

how the publishing industry may be affected by the transition to a liquid publication model. Specifically, we analyse current intellectual property rights schemes and the industry value chain in the current publication model. We then identify innovative services that publishers can offer to the scientific community, to add value and hence generate business, for them.

The discussion of Open Access initiative and self-archiving goes across the four sections, to make them self-contained.

Part 2

Modelling evolutionary, collaborative, multi-faceted knowledge artefacts

A knowledge artefact is an object created as a result of an activity which encodes knowledge or the understanding gained beyond data. Examples of these knowledge artefacts include books, journals, scientific papers, among others.

The adjectives identified at the title of this part refer to desirable qualities to these knowledge artefacts and the following list will give a quick introduction to each of them:

- Evolutionary: the possibility of reuse and evolution is mainly used to help deal with the overhead involved in refining an idea through several iterations. Non-evolutionary methods, on the other hand, require that a completely new artefact is created for even small progress or changes to be introduced. As such, evolutionary knowledge artefacts help the early circulation of innovative ideas, since subsequent iterations become easier to create and manage.

- Collaborative: collaborative artefacts can enable the collaboration and contribution of a number of interested parties. This collaboration can be direct, in which two or more parties work together, relatively at the same time, to a common goal, or take a more indirect approach in which a second party takes a previously developed work as a base for new developments. In all cases, besides the simple addition of work, collaboration enables the coming together of different perspectives, opinions, or interests. Which, along with other properties discussed at this list, may motivate some parties to pursue slightly different approaches making the artefact evolve in a tree-like branched fashion.

- Multi-faceted and multi-purpose: it is not uncommon for a single artefact to be involved on several topics and try to cover different objectives. As such, an explicit support for this variety of content (for example images, videos, datasets) and subjects (for example, biology, computer science, sociology) can make these complex artefacts and their components much easier to find by the possibly interested parties.

- Composite and composable: composing may take place at the idea level, in which several ideas can be composed to form the basis to a new idea or also at the whole artefact level in which it creates collections or compilations. In both cases, having some sort of support

and guide for composition would greatly optimize the time for the creation of composition-based artefacts. This is specially true in artefacts which involve collections of other works like for example, books, journals and conferences.

The rest of this part will describe the current and upcoming approaches and advances in the modelling, accessing, sharing and dissemination of knowledge artefacts. In particular Section 2.1 and Section 2.2 will introduce Scientific and Software artefacts and the similarities and relations existing between them. Then, Section 2.3 will detail the approaches for representing and managing data, knowledge and metadata attributes. Finally, Sections 2.4, 2.5 and 2.6 will deal with the social aspects of artefact creation, with the evolution, tracking and processes of artefacts and with their publishing and dissemination.

2.1 Current Scientific and Research Artefacts

Scientific research in general is described in [43, 169] as having the following properties:

- Monotonic: once something is released and published it cannot longer be unpublished or undone. Later works may be used to expand, extend, modify or contradict previous existing work but the original work always remains.
- Concurrent: several different scientific groups or organizations may work on the same subject at the same time, probably overlapping and duplicating efforts or interacting to attempt to improve each other's performance.
- Plural: publications and the community may have heterogeneous, varied and even conflicting information and opinions. Even so, there is no agreed arbiter to mediate in such conflicts, as the community normally takes upon on itself to solve them.
- Sceptic: publications and researches normally assume that their target is "sceptic by default", so they try to validate their current information and proposed knowledge as much as possible.
- Provenance tracked: the origin of information is strictly tracked and normally referred to by the use of the so called citations or references.
- Commutativity: no officially recognized "given order" is established for having access to publications topics or a community. As such, authors normally try to make the publications as much "self-contained" as possible so they may be read regardless whether they initiate new research or are relevant to already ongoing research.

We will use the name scientific artefacts to denote artefacts that convey scientific data and knowledge. The most common of these is the scientific paper or article, which was introduced during the 17th century when the first academic journals appeared.

Currently, as detailed in [84] and not very unlike those early times, when an author wants to publish a scientific paper he has to submit a physical and or digital copy of it to an academic journal, where it goes through a process of peer reviewing to determine if its publication is suitable.

Scientific journals fulfil then the double role, of certifying (the peer review and eventual acceptance in the journal) and of dissemination (the actual publishing and making it available and accessible to others) of knowledge. The dissemination part also includes the actual archiving and preservation of the knowledge to warrant perennial access.

This model has remained mostly undisturbed up to now, even with the transition to the electronic era reducing the costs implied in the dissemination process and the advent of the Internet and the Web providing new ways of contact and interaction.

As the main issues and disadvantages of the current paper-based publishing and dissemination, works like [68] and [52] agree on the following:

- Delayed dissemination: several months could go by between the submitting of a paper and its publication. This high dissemination overhead is a widely shared concern between researchers, as the delays it causes may hinder collaboration between the groups working independently on similar subjects unless additional publishing options (like making it the article/paper available online) are added on top of the regular publication.
- Too much time spent writing instead of researching: as papers are the main mean for career advancement, too much time is spent writing papers instead of developing actual research.
- Current metrics encourage practices not useful for the progress of science: in particular, the citation count ranking in which authors are rated by the amount of accepted papers they have encourages:
  - superficial research
  - overly large groups (professors with many students who do not spend a lot of time with each)
  - repetition (copy, paste, disguise)
  - small insignificant studies and half-baked ideas
  - amount of papers is more important than correctness, novelty and importance
- Collaboration and composition and incremental contributions not favoured: relying on previously existing work is only enabled by the reference and quotation systems, without additional composition approaches being widely used. Additionally, the requirement of writing full papers normally hampers small incremental changes that may nonetheless be interesting to the community.
- Creation of 'virtuous circles' and 'natural barriers': by top rated authors and journals adjusted to the workings of the current system. This makes competition and introduction of new authors and research groups much harder.
- Flawed review process: the peer to peer review model is known to reject good papers that later go on to win awards of excellency and to approve papers that are later found out to be inaccurate or incorrect [220].
- Processes are unnecessarily joined: scientific knowledge evaluation, dissemination and accreditation are currently all joined and set to a strict order under the paper submission and publication process. Nevertheless, other types of knowledge separate these processes and

even change the order between them, resulting in more agile production and dissemination processes. For example, the blogs are disseminated previous to any review and after the dissemination the community rates and leaves their comments on them.

2.2 Software Artefacts

Software artefacts refer to artefacts that contain encoded instructions for the use of a computer, thus despite also having the possibility of encoding knowledge they are not directly targeted at humans.

The rest of this section will deal with key factors and ideas from the software world that are considered interesting models for improving the creation and distribution of other knowledge artefacts.

2.2.1 Software Development

One of the key insights of this project is that the handling of scientific knowledge resembles the handling of software development. For example, both cases involve collaborative work and exchange of ideas among project members and both normally need several iterations until they reach a final version. However, as a particular difference between the two, it is claimed that the modelling and distribution system for knowledge artefacts has lagged behind the ones from other similar objects, like for example software artefacts.

As an example of the previous, knowledge artefact creation is currently mostly based on the waterfall model, which makes it difficult to adapt to the changes and requirements that are frequently present in the research they represent. While, on the other hand, several other creation, development and distribution models are available and widely used for software artefacts.

One of these creation models such as extreme programming [163] and other agile software process models could be applied to the production management of scientific knowledge as an attempt to bridge several of the delay and collaboration issues existing in the current approach.

2.2.2 Version control systems

Also called Versioning or Revision Control Systems (RCS), these refer to the management of multiple revisions of the same unit of information.RCSs are normally used to enable coordination of work between a group of people or manage the contents or the overall structure [290] resulting from the association of artefacts.

Artefacts are normally continuously rewritten and updated as new ideas and contributions are included in them. This process takes the form of a sequence of iterations applied to the artefact, forming a chain of transitions that contains the evolution of such artefact from its first draft to its final version. Some of these transitions could involve, and update or completion of a previous existing part, recovering ideas from past versions or even the merging of different contributions.

Furthermore, RCS has also been used on non-software artefacts like text editors [60] and are commonly present, in some form, on all the major document edition software.

The following are the major components that form part of RCSs:

- Code repository: a dedicated place where the software stores the entire project code base. This data could reside on the same machine where the development takes place, or on a remote system, accessed via network connections.
- Files: RCSs are about files and their evolution. As such, RCSs need to be able to handle and tell the differences between any kind of file.
- Patches: also called change-sets contain the differences between two versions and also the way of converting the earlier version into the newer one.
- Locking: used to manage the access to the files under version control and enable concurrent access to them and ensure their consistency.

A centralized RCS used for development works perfectly if it is assumed that all the interested users can contact the server at any given time. However this centralization also introduces a single point of failure, as any problems with the central server potentially denies its services to all users at the same time even if these users are somehow able to communicate directly.

As a response to this Distributed Revision Control Systems (DRCSs) were, somewhat recently, introduced. As their main feature DRCSs do not rely on a central repository holding the whole code-base but instead provides each developer with his/her own local repository.

Though this allows each user a great deal of autonomy, as they can branch and fork and update their copy as they see fit, it also introduces some additional considerations about the concurrency and consistency of the repository [145]. Well-known examples of DRCSs include Bazaar[1] Mercurial[2] and Git[3].

Summing up, revision control systems try to tackle two fundamental problems which affect every software development project:

1. Let different developers work on the same project while maintaining one single, coherent code base, thus avoiding the project code to become an inconspicuous mix of reciprocally incompatible contributions;
2. Keep track of every change made to the code base by developers, thus allowing editors to roll back whatever changes they did to the code. This also means that every change bears the name of her/his author, as well as the timestamp on which it was made.

2.2.3 Online Software Development Services

DRCSs also enable large scale and asynchronous coordination of collaboration within a group which, combined with easy of access and openness of the web environment is used to create development services communities that develop software using permissive and open source [158] licenses. Popular on-line open source development :

[1] http://bazaar-vcs.org/
[2] www.selenic.com/mercurial/
[3] http://git.or.cz/

- SourceForge[4], which has been targeted by several studies like [253].
- Ohloh[5], which auto-crawls the projects' repository to present activity and other data.
- Launchpad[6], which is currently used by the Ubuntu Linux project among others.
- Google code[7], which contains open source code based on the Google public APIs for developers interested in Google-related development.

2.3 Data and Knowledge Management

The terms data, information and knowledge are frequently used for overlapping concepts. Data is the lowest level of abstraction, which is the signal without interpretation and processing we touch by the minute. Information is the next level, which is endowed by meaningful data. And finally, knowledge is the highest level among all three that can always be used to generate new information. For knowledge management, there are three main research areas including knowledge retrieval, knowledge representation and knowledge application. This section will introduce a number of prevalent theories and applications related to data and knowledge management and, more specifically, the development of semantic-based technologies.

2.3.1 Metadata

Metadata [200] is "data about data" or "information about data", which is used to facilitate the understanding, characteristics, and management usage of data. For instance, metadata would document data about data elements or attributes (name, size, data type, etc.), and data about records or data structures (length, fields, columns, etc) and data about data (where it is located, how it is associated, ownership, etc.). Metadata may also consist of descriptive information about the context, quality and condition, or characteristics of the data. A metadata set consists of elements in which each element refers to a label that describes particular information about a resource. A resource here is defined as "anything that has an identity". In the context of the Document-like Information Objects, a resource could be, for example, a document, monograph, web page, project report, or even people.

There are several initiatives and applications which aim to encompass issues of semantic standards in the domain of knowledge transaction with respect to description, resource discovery, interoperability and metadata exchange for different types of information resources.

- Dublin Core Metadata Initiative[8] (DCMI) - The Dublin Core Metadata Initiative is an open organization engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models. DCMI's activities include work on architecture and modelling, discussions and collaborative work in DCMI Communities and

[4] http://sourceforge.net/
[5] http://www.ohloh.net/
[6] http://launchpad.net/
[7] http://code.google.com/
[8] Dublin Core Initiative. http://dublincore.org

DCMI Task Groups, annual conferences and workshops, standards liaison, and educational efforts to promote widespread acceptance of metadata standards and practices.

- Friend of a Friend[9] (FOAF) - The Friend of a Friend (FOAF) project is creating a Web of machine-readable pages describing people, the links between them and the things they create and do; it is a contribution to the linked information system known as theWeb. FOAF defines an open, decentralized technology for connecting social Web sites, and the people they describe. Open Archives Initiative Protocol for Metadata Harvesting[10] (OAI-PMH) - It defines a mechanism for data providers to expose their metadata. This protocol suggests that individual archives map their metadata to the Dublin Core, a simple and common metadata set for this purpose.

2.3.2 Ontology

The term ontology [217] has its origin in philosophy. In computer science and information science, an ontology is a formal representation of a set of concepts within a domain and the relationships between those concepts. It is used to define and model the domain knowledge and reason about the properties of that domain.

In theory, an ontology is a ”formal, explicit specification of a shared conceptualisation” [288]. An ontology provides a shared vocabulary, which can be used to model a domain - that is, the type of objects and/or concepts that exist, and their properties and relations [90].

2.3.3 Hypertext

Hypertext is text which is displayed on a computer with hyperlinks[11] to other text or objects that the reader can access directly by a mouse click. It may contain not only text, but also tables, pictures, and other multimedia materials. Hypertext documents can be either static or dynamic. Herein, 'static' means the hyperlinks and the linked data are stored and prepared in advance, while the 'dynamic' one provides a dynamic response according to user's input and request. Nowadays, static hypertext is widely used for cross-reference collections of data in scientific papers and books.

A lot of work has been done in 1980's and 1990's aiming at content linking and reuse. In 1980, Tim Berners Lee created the earliest hypertext database system 'ENQUIRE'. Later, he invented the World Wide Web to meet the demand for document sharing among scientists working in different places all over the world. In 1992, the first Internet browser 'Lynx' was born. It provided hypertext links within documents which could reach into documents anywhere on the Internet. HyperCard was one of the most famous hypertext systems in 1980's and 1990's. From the year of 1987 to 1992, HyperCard was sold with Apple computer as promotions. Thereafter, DHM(Aarhus University, Denmark), Chimera(University of California, US), and Microcosm(Southampton University, UK) became dominant open hypertext system in those days.

In hypertext systems, a typed link is a link to another document or a document part that also

[9] FOAF Initiative. `http://www.foaf-project.org`

[10] Open Archives Initiative Protocol for Metadata Harvesting. `http://www.openarchives.org/OAI/2.0/guidelines.htm`

[11] http://en.wikipedia.org/wiki/Hyperlinks

carries the information about the link type. For instance, rather than merely pointing to the existence of a document, a typed link might also specify the relationship between subject and object, such as 'next' relation, 'previous' relation and so on. This facilitates a user to take actions like searching certain types of links or displaying them differently. In 'HTML 4.01 Specification' [27], W3C predefined a set of link types which are as follows:

- alternate
- stylesheet
- start
- next
- prev
- contents
- index
- glossary
- copyright
- chapter
- section
- subsection
- appendix
- help
- bookmark

SGML(Standard Generalized Markup Language) is development of hypertext technology. It is also an ISO standard aiming to define generalized markup languages for documents. HTML and XML are both derivatives of SGML, while XML is a subset of SGML and HTML is an application of SGML. Some work has been done on content reuse, validation and source attribution as SGML/XML applications especially in the publishing area. EDGAR[12] (Electronic Data-Gathering, Analysis, and Retrieval) system provides automatic collection, validation, indexing, acceptance, and forwarding of submissions. Authorized companies and other organizations can use it to require file data and information forms from the US Securities and Exchange Commission (SEC). DocBook[13] is a semantic markup language originally invented as an SGML application, designed for writing technical documentation related to computer software and hardware. It currently is an XML application. DocBook helps users create and reuse document content in a presentation-neutral form that can capture the logical structure of the document. After that, the

[12]EDGAR: http://www.sec.gov/edgar.shtml
[13]DocBook: http://www.docbook.org/

content can be published in various formats, including HTML, XHTML, EPUB, PDF, etc., without asking users to make any changes to the source attribution.

In general, Hypertext research promotes content reuse, validation, and source attribution. Thus, the definition of a scripting language for authoring, annotating, search, and identification of types and patterns from scientific publications would be a very interesting objective for the LiquidPub project.

2.3.4 Markup Language

A markup language [196] is an artificial language using a set of annotations to describe the information regarding the structure of text or how it is to be displayed, which has been popularly used in computer typesetting and word-processing systems. Within a markup language, there is metadata, markup and data content. The metadata describes characteristics about the data, while the markup identifies the specific type of data content and acts as a container for that document instance. A well-known example of a markup language in use today in metadata processing is HTML, one of the most used in the World Wide Web. Furthermore, The Extensible Markup Language (XML) [303, 314] is a general-purpose specification for creating custom markup languages. It is classified as an extensible language, because it allows the user to define the markup elements. XML's purpose is to aid information systems in sharing structured data, especially via the Internet, to encode documents, and to serialize data. Similarly, LaTeX is also a well-know example of a markup language but used mainly within the domain of documents and typesetting.

The Resource Description Framework (RDF) [237] is a family of World Wide Web Consortium (W3C) specifications [236], originally designed as a metadata data model. It has come to be used as a general method for conceptual description or modelling of information that is implemented in web resources; using a variety of syntax formats.

The Web Ontology Language (OWL) [307] is a family of knowledge representation languages for authoring ontologies, and is endorsed by the World Wide Web Consortium. This family of languages is based on two (largely, but not entirely, compatible) semantics: OWL DL and OWL Lite semantics are based on Description Logics [12], which have attractive and well-understood computational properties, while OWL Full uses a novel semantic model intended to provide compatibility with RDF Schema. OWL ontologies are most commonly serialized using RDF/XML syntax. OWL is considered one of the fundamental technologies underpinning the Semantic Web, and has attracted both academic and commercial interest.

Lemon8-XML[14] is an effort of Public Knowledge Project (PKP) led by University of British Columbia and Stanford University. It's designed to help editors and authors convert scientific papers from typical editing formats such as MS-Word and OpenOffice, into XML-based publishing layout formats. It provides the ability to edit document and its metadata as well. Lemon8-XML is a stand-alone system that serves publishing processes more generally.

[14] http://pkp.sfu.ca/lemon8

2.3.5 Formal Classification and Ontology

In today's information society, as the amount of information grows larger, it becomes essential to develop efficient ways to summarize and navigate information from large, multivariate data sets. The field of classification supports these tasks, as it investigates how sets of "objects" can be summarized into a small number of classes, and it also provides methods to assist the search of such "objects" [2]. In the past centuries, classification has been the domain of librarians and archivists. Lately a lot of interest has focused also on the management of the information present in the web: see for instance the WWW Virtual Library project[15], or directories of search engines like Google, or Yahoo!

Standard classification methodologies amount to manually organizing topics into hierarchies. Hierarchical library classification systems, such as the Dewey Decimal Classification System (DDC)[16] or the Library of Congress classification system (LCC)[17], are attempts to develop static, hierarchical classification structures into which all of human knowledge can be classified. ACM (Association for Computing Machinery) has published a computing classification system[18] for the computing field. Although these are standard and universal techniques, they have a number of limitations. Classifications describe their contents using natural language labels, an approach which has proved very effective in manual classification. However natural language labels show their limitations when one tries to automate the process, as they make it almost impossible to reason about classifications and their contents.

An ontology is a formal representation of a set of concepts within a domain and the relationships between those concepts. It is used to reason about the attributes or metadata of that domain, and may be used to define the domain. Ontologies are used in artificial intelligence, the Semantic Web, software engineering, biomedical informatics, library science, and information architecture as a form of knowledge representation about the world or some part of it.

When dealing with classifications, an innovative method [104] is to formalize and encode the classifications into lightweight ontologies [105]. And then, to reason about them, to associate to each node a normal form formula which uniquely describes its contents, and to reduce document classification and query answering to reasoning about subsumption.

When dealing with classifications, an innovative method [104] is to formalize and encode the classifications into lightweight ontologies [105]. And then, to reason about them, to associate to each node a normal form formula which uniquely describes its contents, and to reduce document classification and query answering to reasoning about subsumption. In the report by Olena Medelyan et al [199], some novel methods to extract concepts and relations from Wikipedia[19] for ontology building are discussed.

[15] `http://vlib.org/`
[16] `http://www.oclc.org/dewey/`
[17] `http://www.loc.gov/catdir/cpso/lcco/lcco.html/`
[18] `http://www.acm.org/about/class/`
[19] http://en.wikipedia.org

2.3.6 Semantic Matching

Semantic matching is a technique used in Computer Science to identify information which is semantically related. Given any two graph-like structures, e.g. classifications, database or XML schemas and ontologies, matching is an operator which identifies those nodes in the two structures which semantically correspond to one another. For example, applied to file systems it can identify that a folder labelled ”car” is semantically equivalent to another folder ”automobile” because they are synonyms in English. This information can be taken from a linguistic resource like WordNet.

In the recent years many techniques has been offered. A good survey is represented by [110]. S-Match [87, 105, 107, 106, 109, 108] is a good example of a semantic matching operator. It works on lightweight ontologies, namely graph structures where each node is labelled by a natural language sentence, for example in English. These sentences are translated into a formal logic formula (according to an artificial unambiguous language) codifying the meaning of the node taking into account its position in the graph. For example, in case the folder ”car” is under another folder ”red” we can say that the meaning of the folder ”car” is ”red car” in this case. This is translated into the logical formula ”red AND car”.

The output of S-Match is a set of semantic correspondences called mappings attached with one of the following semantic relations: disjointness, equivalence, more specific and less specific. In our example the algorithm will return a mapping between ”car” and ”automobile” attached with an equivalence relation.

Semantic matching represents a fundamental technique in many applications in areas such as resource discovery, data integration, data migration, query translation, peer to peer networks, agent communication, schema and ontology merging. In fact, it has been proposed as a valid solution to the semantic heterogeneity problem, namely managing the diversity in knowledge. Interoperability among people of different cultures and languages, having different viewpoints and using different terminology has always been a huge problem. Especially with the advent of the Web and the consequential information explosion, the problem seems to be emphasized. People face the concrete problem to retrieve, disambiguate and integrate information coming from a wide variety of sources.

2.3.7 Semantic Search

The goal of Semantic Search [255] is to augment and improve the search process by leveraging XML and RDF data from semantic web to produce highly relevant results. The key distinction between semantic search and traditional search is that semantic search is based on semantics, while traditional one mainly focuses on keywords mapping. Guha et al. [230] distinguished two major kinds of search: navigation and research. In navigational search, the user is using the search engine as a navigation tool to navigate to a particular intended document. Semantic Search is not applicable to navigational searches. In research search, the user provides the search engine with a phrase which is intended to denote an object about which the user is trying to gather. There is no particular document which the user knows explicitly about that s/he is trying to get to. Rather, the user is trying to locate a number of documents which together will give him/her the information s/he is trying to find Semantic Search lends itself well here.

Hildebrand et al. [256] provide a survey that lists semantic search systems and identifies

other uses of semantics in the search process. Besides, W3C maintains a list of Semantic Web Tools [257] and various others exists, such as the Developers Guide to Semantic Web Toolkits and the Comprehensive Listing of Semantic Web and Related Tools by Michael K. Bergman[20]. These lists support the semantic web community (mainly developers) by providing an overview of available tools. We introduce here several well-known online semantic search engines specialized in OWL and/or RDF content.

Falcon[21] is a keyword-based search engine for the Semantic Web, equipped with browsing capability. Falcons provides keyword-based search for URIs identifying objects, concepts (classes and properties), and documents on the Semantic Web. Falcons provides a summary for each entity (object, class, property), integrated from all over the Semantic Web [316, 116, 117]. It also attempts to classify a search phrase by type of search, e.g. Person, Document.

Sindice[22] is a lookup index for Semantic Web documents. Sindice indexes the Semantic Web and can tell users which sources mention a resource URI, IFP, or keyword. Sindice does not answer triple queries. Users can use Sindice in their application to find relevant RDF sources. Over 10 billion pieces of reusable information has already been indexed across 100 million web pages which embed RDF and Microformats in Sindice.

Swoogle[23] is a search engine for the Semantic Web on the Web. Swoogle crawl the World Wide Web for a special class of web documents called Semantic Web documents, which are written in RDF. Currently, it provides the following services to the following services:

- search Semantic Web ontologies
- search Semantic Web instance data
- search Semantic Web terms, i.e., URIs that have been defined as classes and properties provide metadata of Semantic Web documents and support browsing the Semantic Web [184]
- archive different versions of Semantic Web documents

Currently, Swoogle only indexes some metadata about Semantic Web documents. It neither stores nor searches all triples in a Semantic Web documents as a triple store.

SWSE (Semantic Web Search Engine)[24] provides an entity-centric view on Semantic Web instances. SWSE also provides a SPARQL (SPARQL Protocol and RDF Query Language)[25] endpoint over currently around 400k RDF files from the Web. SWSE attempts to offer a service which continuously explores and indexes the Semantic Web and provides an easy-to-use interface through which users can find the data they are looking for.

[20] http://www.mkbergman.com/?p=291
[21] Falcon. `http://iws.seu.edu.cn/services/falcons/conceptsearch/`
[22] Sindice – The Semantic Web Index. `http://sindice.com/`
[23] Swoogle – Semantic Web Search. `http://swoogle.umbc.edu/`
[24] SWSE – Semantic Web Search Engine. `http://swse.org/`
[25] SPARQL Query Language for RDF. `http://www.w3.org/TR/rdf-sparql-query/`

2.3.8 Document Management

Traditional document repositories based on hierarchical organizations such as filesystems, offer a limited view of a document space. Bush envisioned in 1945, some of the principles of that apply to modern document management systems; in which annotations, tagging and search are important features. In his work [45], Bush motivates the use of associations in his hypothetical device memex instead of indexes as they are closer to the way human mind works. Some tools were developed to fulfil the vision behind memex [99].

The placeless document management system [76], or Placeless for short, is based on document properties rather than location in order to provide document organization. It extends from the normal document properties (i.e., creator, creation date, etc.) to offer universal properties (related to the base document) and personal properties (associated to document references). A key point of this proposal is the possibility of associating functionality to documents by means of active properties (i.e., properties with code associated). Thus, documents can raise actions when these properties are set (e.g., setting collaborator='John Doe' can send a mail to the collaborator). Another important feature is the separation of content (physical data) from the document management. This allows reusing existing repositories and document processing tools.

Defining document workflows or lifecycles is another aspect of document systems. In some cases, the processes involving the document evolution are pre-defined (for example, a document system with focus on digitalization). In most advanced systems instead, a workflow or lifecycle facility is provided.

A document workflow approach based on Placeless is introduced in [173]. In this, active properties are used to provide the workflow functionality. Moreover, as with most real world scenarios, flexibility is also important in this context. A framework for document-driven workflows was proposed in [148], which requires no explicit control flow. In this approach, the boundary of the flexibility is described by the dependency among documents (in terms of inputs and outputs). Nevertheless, as workflow operations are associated to changes in the documents, these changes must be done under the control of the workflow.

In a more general approach introduced in [1], the processing of artefacts, from the creation to completion and archiving, is captured by lifecycles. Nonetheless, the flexibility offered is more focused on the artefact representation rather than lifecycle evolution and execution.

Besides providing organization, annotations and search capabilities and workflow management; document management systems provide usually versioning, sharing, document processing, and publishing.

A new scenario arises as new sophisticated tools are developed using the Web 2.0 model. Services such as Google Docs, Zoho Docs and social sharing systems are gaining more and more users due to their sharing and collaboration features. This definitely imposes new requirements as data becomes disperse, personal workspaces intersect with organizational workspaces, new devices and publishing platforms become available. Thus, interesting conceptual and design problems emerge in this ongoing area of research.

2.4 Collaborative Approaches

This section discuses the role of the ICT technologies in bridging the distances and enabling the collaboration of the creation and evolution of artefacts.

2.4.1 Online Social Networks

Social web services are based on communities of people that are brought together by the use of services like e-mail, forums among others. Enabled by these services social networks are formed [14] and, as these networks continue to grow in specific ways and patterns [233], their users normally collaborate in the creation of several types of artefacts among other activities. Artefacts created within social networks include the following particularities:

- artefacts over the social networks are created/modified and distributed at the same time (in fact the one is directly related to the other), which is a notable change from previous approaches.
- In most of the cases there is not a fixed end for the artefact's evolution, which is beneficial in the sense that it keeps the artefact updated but it may also cause its accuracy to be compromised.
- A relatively large amount of people has access to commenting, reviewing and sometimes even directly modifying the artefact.
- Having this large number of individuals collaborate is the key detail of the approach. This collaboration normally does not take into consideration the credentials of the persons and the sheer volume people collaborating can make the credit attribution hard to do accurately.
- Offers additional services like live references, classificatory tools, functionalities that allow the composition of parts and sections, topic-based recommendations, among other.

Depending on the nature of the social network the motivation for the creation of these artefacts also varies. On more communication-oriented networks like [26] users produce artefacts like photo albums to characterize people or places. On the other more content-oriented social networks like Wikipedia[27] focus on the creation of knowledge artefacts through collaboration of a very large number of persons. While Wikipedia has been criticized because its model favours consensus and popular information over accuracy and its restriction of dealing only with well-known subjects (which disallows research), other similar approaches like Swiki[28], Knol[29] and Ylvi[30], introduce modifications to Wikipedia's formula to deal with these limitations.

The following examples present interesting approaches to giving access, disseminating and interacting with knowledge artefacts through a social network:

[26] http://www.facebook.com/
[27] www.wikipedia.com
[28] http://wiki.squeak.org/swiki/
[29] knol.google.com
[30] http://www.cs.univie.ac.at/project.php?pid=268

- *PLos One*[31]: Open Access journal in which submissions go normally through a rigorous(no editor) peer review before publication. One of the key factors in Plos One is however that considers post-publication (commenting and discussing online) as important as the prepublication of its submissions [322].
- *Faculty of 1000*[32]: website for post publication discussion, ranking and commentary on the current research papers and fields. The Faculty of 1000 runs on the same principles of the Web 2.0, meaning that the community's contributions by filtering, tagging, reviewing, ranking, etc. are considered an integral part of the platform [160].

More examples of social networks, Web 2.0 and their implications will be discussed in the following chapters.

2.4.2 Computer Supported Cooperative Work

While the previously discussed social web applies in general to all activities including ludic or other social interactions, social software within the context of work or creative processes is generally known as Groupware, or Collaborative Software.

Groupware forms the basis of Computer Supported Cooperative Work (CSCW), whose main objective is to study the technology to support people in their work [121] addressing how collaborative activities and their coordination can be supported by means of computer systems [51]. A common way of organizing the CSCW is to consider the work context that needs to be dealt with along two dimensions: space and time. This is the so-called CSCW matrix which identifies 4 distinct areas:

- *Same time and same place*: meeting rooms, displays and other face to face tools.
- *Different time and same place*: office rooms, large public displays or post its.
- *Same time and different place*: instant messaging [209], video conference software, etc.
- *Different time and different place*: email, blogs, forums, and other communication/coordination tools.

As a result CSCW studies several software tools like Collaborative editors [75] and the concurrent work and use of resources by humans [119] that are significant for our current study. For more information a Handbook about CSCW, including the top cited 47 papers regarding it, can be found in [201].

2.4.3 Google Wave

Google Wave[33] is a new communication and collaboration tool on the web, coming later this year. Announced on July 20, 2009, it is expected that an official preview will be made available on

[31] www.plosone.org
[32] www.facultyof1000.com
[33] http://wave.google.com/

September 30, 2009, while the number of testing user will be expanded to 100,000 by that time. A wave is equal parts conversation and document. People can use it to communicate and work together with different formatted text, videos, photos, maps and so on. Another feature of a wave is "shared". Any participant can take part in conversations, invite other people, and edit the message together at any point in the process. Especially they can playback the conversation helping people rewind the wave to see who said what and when. Moreover, a wave is live in real-time. Participants can see other's edits in real-time, word by word, and enjoy faster conversations. Google Wave integrates Web-based services, computing platforms and communication agreements, designed to consolidate the e-mail, instant messaging, wiki and social networking. There are some key technologies applied in Google Wave, such as real-time collaboration, natural language tools and extending Google Wave, which provide powerful real-time collaboration, effective spell checking that can automatically translate 40 languages, and many other extensions.

2.5 Artefact-Centred Lifecycles

Most of the real world scenarios are dynamic and thus trying to anticipate a process model is impractical or too complex. Even so, current workflows systems are fairly rigid and prescriptive, and therefore not suitable for dynamic environments. Flexible workflows make a step forward by allowing flexibility in the process modelling and enactment. Nonetheless, when it comes to defining the evolution of artefacts, the artefact-centred approach is the most suited paradigm.

In the artefact-centred approach, the artefact evolution is captured by lifecycles. A lifecycle defines a course of stages in the artefact evolution, depicting a progression (or series of transformations) towards a specific goal (or product). Examples of this are: a deliverable elaboration lifecycle, a paper publication lifecycle, etc.

In our discussion, we take as reference points the characteristics of the scenario considered in LiquidPub. In this, an ideal approach should encourage reusability, abstraction and evolution of lifecycle models, while keeping the model simple and flexible.

This summary presents the relevant literature in three different areas related to this approach: workflow management systems, document management and lifecycle notations. For these, we draw the characteristics that relate them to our work, taking the ideal case as point of comparison.

2.5.1 Workflow Management Systems

Workflow systems [73] allow the definition, execution, and management of workflows. In general, workflow systems describe a business process as a set of tasks, to be executed in the order defined by the model. The main goal of a workflow system is that of process automation, both for human-oriented processes (in which case the workflow automates the scheduling of tasks to be executed and the transfer of information among human agents) and for production processes (in which case the workflow system is a middleware automating the flow of information among applications). With the advent of Web services, workflow models, languages, and tools evolved to be able to integrate services, by allowing the definition of composite services (services developed by composing other services based on a process logic, and themselves exposed as services) [5].

Workflow systems are related to artefact-centred approach since they describe a flow model

and actions to be executed on objects. They are however different since:

- they do not focus on lifecycle management. They do not focus on the evolution of an object, but rather they model arbitrary actions to be executed by human or automated resources. A lifecycle instead, model the phases in the evolution of an object;
- they are fairly rigid and prescriptive (they work well for structured, repeatable processes)
- they are targeted to programmers and often designed for mission-critical applications (in fact they are not significantly less complex than Java, for example)
- the corresponding software platform is large and complex to operate and maintain

Interesting lessons can however be learned by looking both at research in workflow evolution [85, 86] and adaptive workflow [190, 224, 313] and at research on semi-structured workflow models, including in particular scientific workflows [191, 11] that are targeted at scientists.

An approach to workflow evolution is introduced in [85]. In this, particular attention is put on changes in the workflow specification (static evolution) and their propagation to workflow instances (dynamic evolution). Thus, it allows schema evolution while respecting the structural consistency (legal modification of the workflow schema) and behavioural consistency (transformations that do not result in run-time errors). Nonetheless, the approach does not address issues like deviations from workflow schema during workflow enactment (i.e. skipping an activity) and ad-hoc changes in individual workflow instances. The ideas of schema evolution can be, however, extended to lifecycle schema evolution.

In the area of adaptive workflows, several approaches have been proposed to provide dynamic process management. In these, flexibility is provided by supporting schema evolution and modification in workflow running instances [242, 190, 224]. In [190] Reichert and Dadam build upon a graph-based workflow model (ADEPT) to provide a formal foundation for supporting dynamic changes in running instance workflows. Their approach concentrates on structural changes while preserving consistency and correctness. An extension of this work [224] provides orchestration, allows composing processes from existing application components and supports both ad-hoc deviations and process schema evolution. Nonetheless, these approaches are based on workflow and do not allow to exploit user knowledge and decisions to drive the flow.

Case handling [313] constitutes a data-driven approach to flexible business processes. In this context, case is the product being manufactured, and whose state is the primary driver to determine which activities or unit of works are enabled at a given point in the process. Thus, flexibility is provided by allowing knowledge workers to decide how to accomplish the business goals, that is, which of the available activities to execute. However, this approach does not support run-time model changes. Nevertheless, allowing users to drive the process execution is an interesting way of exploiting knowledge-intensive business processes.

The PROSYT system [65] considers another approach to flexibility in the process enactment. PROSYT takes the artefact-based approach in which operations and conditions for these operations can be defined over the concept of artefact type. Thus, users can deviate from these conditions to execute operations. Nonetheless, each artefact type defines just one possible lifecycle, and runtime lifecycle model changes are not allowed.

With a different target, scientific workflows were developed for scientific problem-solving environments, in which experiments need to be conducted [264, 11]. Experiments can be considered as sets of actions operating on large datasets [191]. Due to the nature of the environment, it is often not possible to anticipate a scientific workflow, so model-changes and user intervention at runtime are necessary to provide flexibility. Other requirements like reproducibility, detailed documenting and analysis are also of concern [264]. Nonetheless, scientific workflows focus on workflows more than lifecycles of artefacts and require a level of expertise which is well beyond the level of average users.

2.5.2 Document Workflow

The artefact-centred approach has roots in the document engineering community [240]. In this area, models and tools are developed around the concept of documents, which can be seen as particular types of artefacts.

In [173] the notion of document-centred collaboration is introduced. There, the activities of collaboration and coordination are considered aspects of the artefact rather than workflows. For this, they attach computation to documents (i.e. a word processor), whose actions define the workflow. However, this approach is focused on decoupling documents from workflows rather than providing a workflow modelling approach.

Flexibility is important in the document management area. A framework for document-driven workflows, which requires no explicit control flow, was proposed in [148]. In this approach, the boundary of the flexibility is described by the dependency among documents (in terms of inputs and outputs). Nevertheless, as workflow operations are associated to changes in the documents, these changes must be done under the control of the workflow, and thus, coupling artefact processing to the workflow. A different approach introduced in [231], addresses the problem of integrating document processing with workflow management system functionalities. In doing so, this approach deals with issues like concurrent access in collaborative environments, which is an aspect of the artefact rather than workflows.

A more general approach, introduced the concept of business artefact, referring to a concrete identifiable, self describing chunk of information that business creates and maintains [1]. In this approach, the processing of artefacts, from the creation to completion and archiving, is captured by lifecycles. Nonetheless, the flexibility offered is more focused on the artefact representation rather than lifecycle evolution and execution.

As described in a recent work [240], some interesting issues to explore in the artefact-centred approach are; i) evolution of the workflow schema, ii) generalization and specialization of workflows, and iii) componentization and composition of workflows. These issues constitute desired qualities for the scenario considered in LiquidPub.

2.5.3 Lifecycle Modelling Notations

At present, there are a variety of models, notations, and languages for describing lifecycles [66, 29, 65], UML being the most popular model [29]. In UML the most common approach is to model lifecycles using state machines that have exactly the purpose of modelling the state and evolution

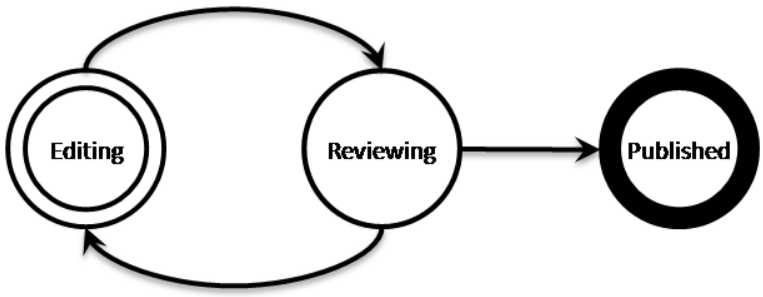

Figure 2.1: Example of a lifecycle modelled with finite state machines.

of an object, and the events that cause state transition [66]. State machines have been extended in a variety of ways, e.g., by allowing guards to be placed on transitions, to associate actions to transitions (state charts [66]), and the like. Figure 2.1 depicts an example of an artefact lifecycle.

Other notations have been used to model lifecycles and processes (see [293] for a survey). The most common ones are Petri nets [312] and activity diagrams [244] and their variations and extensions [293] (which include also workflows and service composition notations such as BPMN [293]). However, these notations are more appropriate for describing workflows and procedures (generic sets of actions to be executed according to some ordering constraints) more than lifecycles (evolution of the state through which a resource goes through, and allowed actions in each state).

2.6 Open Access Initiative

Open Access helps research articles in all academic fields freely available on the Internet to all scientists, scholars, teachers, students, and any other curious mind over the world. It was firstly widely promoted in Budapest in the year of 2002 as an initiative[34]. Nowadays, more and more scholarly journals have successfully supported their free and unrestricted on line availabilities to readers through some dominant web platforms or the websites themselves. Open Access has been perfectly proved its feasibility and efficiency which becomes a popular way of scientific literature dissemination.

[34]Budapest Open Access Initiative. `http://www.soros.org/openaccess/`

2.6.1 Strategies

Peter Suber has proposed a definition of Open Access (OA) in which Open Access literature is digital, on line and with respect of having access to it, free of charge, and free of most copyright and licensing restrictions [275]. The two key factors of Open Access are Internet and consent of author or copyright-holder.

Since authors are not normally paid by scholarly journals, it seems that they are willing to support OA without losing revenue. Besides, OA is totally compatible with peer review while all the major OA initiatives insist on its importance. A significant challenge for OA is the delivery and charge policy to maintain producing although it is less expensive than conventionally published literature. Budapest Open Access Initiative (BOAI) recommended two primary strategies to achieve open access to scholarly journal literature [274].

- OA archives or repositories do not perform peer review, but simply make their contents freely available to the world. They may contain unrefereed preprints, refereed postprints, or both. Archives may belong to institutions, such as universities and laboratories. They could also be classified by disciplines, such as physics and economics. Authors may archive their preprints without anyone else's permission, and a majority of journals[35] already permit authors to archive their postprints. When archives comply with the metadata harvesting protocol of the Open Archives Initiative, then they are interoperable and users can find their contents without knowing which archives exist, where they are located, or what they contain. There is now open-source software for building and maintaining OAI-compliant archives and worldwide momentum for using it.

- OA journals perform peer review and then make the approved contents freely available to the world. Their expenses consist of peer review, manuscript preparation, and server space. OA journals pay their bills very much the way broadcast television and radio stations do: those with an interest in disseminating the content pay the production costs upfront so that access can be free of charge for everyone with the right equipment. Sometimes this means that journals have a subsidy from the hosting university or professional society. Sometimes it means that journals charge a processing fee on accepted articles, to be paid by the author or the author's sponsor (employer, funding agency). OA journals that charge processing fees usually waive them in cases of economic hardship. OA journals with institutional subsidies tend to charge no processing fees. OA journals can get by on lower subsidies or fees if they have income from other publications, advertising, priced add-ons, or auxiliary services. Some institutions and consortia arrange fee discounts. Some OA publishers waive the fee for all researchers affiliated with institutions that have purchased an annual membership. There is a lot of room for creativity in finding ways to pay the costs of a peer-reviewed OA journal, and we are far from having exhausted our cleverness and imagination.

2.6.2 Applications

OA publishing can be traced back to the Los Alamos national laboratory in the United States, while in 1991 it built a website devoting to post original high-energy physics pre-printed research

[35] http://www.doaj.org/ for example

papers (arXiv.org[36]). After that, BioMedCentral[37] (BMC) and PublicLibraryofScience[38] (PLoS) established with active promotion by academic communities and foundations. A large scale scientific literature full-text search engine CiteSeer[39] is another typical case in OA. We introduce here some more well-known OA web platforms as follows.

- *Directory of Open Access Journals*: DOAJ[40] is founded in May, 2003 by Lund University, Sweden. They aim to cover all subjects and languages. There are now 3824 journals in the directory. Currently 1349 journals are searchable at article level. As of today 249182 articles are included in the DOAJ service.

- *High Wire Press*: HighWire[41], a division of the Stanford University Libraries, hosts the largest repository of high impact, peer-reviewed content, with 1203 journals and 5,105,362 full text articles from over 140 scholarly publishers. HighWire-hosted publishers have collectively made 1,831,482 articles free. With their partner publishers they produce 71 of the 200 most-frequently-cited journals.

- *Japan Science and Technology Agency*: JST[42], is an integrated organization of science and technology in Japan that aims to promote dissemination of scientific and technological information. It provides a database named J-STORE[43] for open access to patents and other research results.

- *The Scientific Electronic Library Online*: SciELO[44] is an electronic library covering a selected collection of Brazilian scientific journals. It envisages the development of a common methodology for the preparation, storage, dissemination and evaluation of scientific literature in electronic format.

- *OAIster*: OAIster[45] currently provides access to 19,499,841 records from 1069 contributors, which is developed by University of Michigan. It is a union catalogue of digital resources. They provide access to these digital resources by "harvesting" their descriptive metadata (records) using OAI-PMH (the Open Archives Initiative Protocol for Metadata Harvesting). OAIster can be searched by Title, Author/Creator, Subject, Language or Entire Record. Searches can also be limited by resource type (text, image, audio, video, dataset) and sorted by title, author, date and hit frequency. Results allow further limiting by data contributor (i.e., where the record was harvested from).

[36] http://arxiv.org/
[37] BioMedCentral - The Open Access Publisher. http://www.biomedcentral.com/
[38] Public Library of Science. http://www.plos.org/
[39] CiteSeer, IST - Scientific Literature Digital Library. http://citeseer.ist.psu.edu/
[40] Directory of Open Access Journals. http://www.doaj.org/
[41] High Wire Press. http://highwire.stanford.edu/
[42] Japan Science and Technology Agency. http://www.jst.go.jp/EN/
[43] http://jstore.jst.go.jp/EN/
[44] The Scientific Electronic Library Online. http://www.scielo.br/
[45] OAIster. http://www.oaister.org/

2.7 Scientific Artifact Platforms and Semantic Web for Research

Semantic Web and Social Networking Services promote the evolution of managing research resources. Nowadays, more and more online portals and software platforms provide features combining these two techniques, i.e. research social networking management with Semantic technologies. HypER[46] , Papyres [208], Cyclades[47],and Mendeley[48] are the most famous web applications for scientific publications' sharing and collaborating by Web 2.0 at both data and metadata levels, while Galaxy Zoo[49], Cohere[50], and CORAAL[51] are more focused on some certain specific procedures for scientific artifact discovering and disseminations, like indexing and search. Besides these, there are several predefined modularities and ontologies for scientific publications and research communities which constitute the foundations of semantic annotation(typing), scientific artifact modularizing analysis(patterning) and argumentation representation. The widely used ones are as follows.

2.7.1 ABCDE Format

The ABCDE Format [8] is proposed by Anita de Waard et al., which provides an open standard and widely reusable format for creating rich semantic structures for the articles during writing. The ABCDE stands for Annotation, Background, Contribution, Discussion, and Entities respectively. Using this format, people can easily mark papers semantically, especially in the LaTex editing environment.

2.7.2 ScholOnto

The Scholarly Ontologies Project[52] led by Simon Buckingham Shum et al. in Open University aims at building and deploying a prototype infrastructure for making scholarly claims about the significance of research documents. 'Claims' are made by building connections between ideas. The connections are grounded in a discourse/argumentation ontology, which facilitates providing services for navigating, visualizing and analysing the network as it grows ([263], [292] and [258]). They also implemented a series of software such as ClaiMaker[53], ClaimFinder[54], ClaimBlogger[55] and so on.

[46] http://hyp-er.wik.is/
[47] http://www.ercim.org/cyclades/
[48] http://www.mendeley.com/
[49] http://www.galaxyzoo.org/
[50] http://cohere.open.ac.uk/
[51] http://coraal.deri.ie:8080/coraal/
[52] http://projects.kmi.open.ac.uk/scholonto/index.html
[53] http://claimaker.open.ac.uk/
[54] http://projects.kmi.open.ac.uk/scholonto/software.html#claimfinder
[55] http://claimblogger.open.ac.uk/

2.7.3 SWRC

The SWRC[56] (Semantic Web for Research Communities) project specifies an ontology for research communities, which describes several entities related to research community like persons, organizations, publications and their relationships [278]. It is widely used in a number of applications and projects such as AIFB portal[57], Bibster[58] and the SemIPort project[59]. It aims at facilitating scientific resources' distribution, maintenance, interlinking and reuse.

Bibliographic Ontology[60] specifies sets of concepts and attributes for describing citations and bibliographic references (i.e. papers, books, etc) on the Semantic Web. It could be used as a bibliography ontology, a document classification ontology, or a common ontology for describing any general document. It is also compatible with many other existing document description metadata formats, like Dublin Core and so on. Zotero[61] is one of the most famous applications of this ontology.

2.7.4 SALT

SALT http://salt.semanticauthoring.org (Semantically Annotated LaTeX) is developed by Digital Enterprise Research Institute (DERI) Galway. It provides a semantic authoring framework which aims at enriching scientific publications with semantic annotations, and could be used both during authoring and post-publication. It consists of three ontologies, i.e. Document Ontology, Rhetorical Ontology, and Annotation Ontology, which deal with annotating linear structure, rhetorical structure, and metadata of the document respectively [120]. The annotation ontology is also an extension and implementation of ABCDE format.

In general, the LiquidPub project aims to draw inspiration from the successes of these ongoing works and propose some additional solutions for the managing of scientific knowledge creation, evolution, collaboration and dissemination.

As an example, it would be interesting to provide viable, simple and intuitive means for the creation of semantic documents for scientific publications. In particular, LiquidPub would aim to define metadata structures for scientific publications and their parts, along with the creation of an user environment where this metadata can be be easily completed or inferred from user actions. Thanks to this, it would become possible to offer comparatively better creation tools for documents and contributions, search and navigation, person and contribution discovery among other features that are the objective of the LiquidPub project.

2.8 Conclusion

The previous sections have surveyed how scientific and other related artefacts are created, evolved, disseminated, shared and accessed through the use of current technology. Several of the desirable

[56] http://ontoware.org/projects/swrc/
[57] http://www.aifb.uni-karlsruhe.de/english
[58] http://bibster.semanticweb.org/
[59] http://km.aifb.uni-karlsruhe.de/projects/semiport
[60] http://bibliontology.com/
[61] http://www.zotero.org/

qualities that the project would like to apply to the artefacts and its processes (evolutionary, collaborative, multi-faceted, among others) where also covered in detail. However, no single approach combining to all the requirements from the project was found between them.

This information will be used, in the context of the project, as the starting base and inspiration for the proposal of a model that would enable artefacts, capable of encoding science and knowledge to incorporate all of those desirable qualities. Furthermore, as detailed in the objectives of Liquid Publications proposal document, through this model implement improvements to the creation, evolution, dissemination, review and accreditation of these artefacts.

Part 3

Analysis of reviews and modelling of reviewer behaviour and peer review

This section is concerned with two topics: efforts to analyse review processes in various scientific fields and existing approaches to modelling reviewer behaviours. Besides peer reviews and reviewers, the study also touches upon how reviews are done in related fields where creative content is produced such as the evaluation of software artefacts, pictures, or movies. We also identify metrics for 'good' reviews and review processes, and how to construct them.

Peer review is simultaneously one of the most entrenched and most controversial aspects of research assessment. Virtually every active researcher has experienced their papers or research proposals being blocked by reviews that seemed quite overtly malicious and perhaps even mendacious. At the same time most of us have also at one time or another gained great benefit from a referee who helped to correct unnoticed (and sometimes serious) errors, suggested ways to clarify or improve our results and the description of them, or brought to our attention other related work that we found of great interest. We live in fear of the first kind of reviewer and hoping to find the second—yet it seems rare that there is any consistency in the process.

Stated opinions about peer review range from it being 'crude and understudied, but indispensable' [159] to 'a flawed process' whose effectiveness is a matter of faith rather than evidence [267]. Accusations [20] include systematic bias on the grounds of gender, status and other issues, inconsistency of results, inhibition of innovation, and sheer ineffectiveness, particularly when faced with fraud. Deliberate abuse is also often suspected [25, 177], with reviewers using the process to block the work of rivals or scientists they do not like, to take revenge for rejections of their own work, to censor opinions they dislike and sometimes even to steal results or ideas[1].

Overall, perhaps the most striking thing to be found when examining peer review and the various studies performed on it is the sheer lack of agreement on every single aspect—from whether or not certain problems exist to the most basic question of what peer review is for and how it should be done [155, 267]. Meta-studies are often unable to pool the results of individual investigations

[1] Smith [267] cites a rather shocking case reported by New England Journal of Medicine editor Drummond Rennie, where a reviewer, having produced a critical report on a submitted manuscript, then copied several paragraphs and submitted this 'new' work to another journal. He was found out when his own manuscript was sent for review to the original author. Nature experienced a similar situation where a referee held up the publication of a paper while using his privileged position to obtain materials to assist his own work and so scoop the original author [126].

because processes of review (and investigation of those process) vary so greatly in design [154].

It follows that in some respects this 'state of the art' report presents a rather uncomfortable picture: that in the present state, there seems to be very little art indeed. On the other hand, one absolutely consistent theme of the literature on peer review is the question, 'If not peer review, then what?' Demonstrably effective alternatives to the current system(s) are therefore highly desirable, and with the new technologies and distribution methods offered by the Internet, a variety of new techniques become available. We also examine review and quality assurance techniques from outside academia, notably from the free and open source software communities [235] and new community-created reference works such as Wikipedia and Citizendium.

3.1 Analysis of peer review

Though we may speak of 'peer review' in a single breath, the term actually covers a wide range of activities and conventions that have arisen at different times and places for often very different reasons. In fields such as maths and physics that have a strong preprint tradition, work is frequently shared, used and cited long in advance of its peer-reviewed publication. On the other hand in fields such as biomedical research, where there are strong considerations related to human health, ethical practices and commercial interests, practices are much more stringent and a work may be deemed to not even exist until it has been published in a peer-reviewed journal. Understanding peer review, its practices, benefits and flaws, therefore means an appreciation of both the professional and historical context and active empirical study of the real (as opposed to intended) consequences of different peer review practices. In this section, therefore, we review both the history of peer review in scholarly publishing, and the increasingly large body of work that has been done to investigate the practice and effects of the peer review process.

3.1.1 History and practice

Review by peers in one form or another has been a method of evaluation since Greek times [13, 269] and has been a formal part of scientific communication since the first scientific journals appeared over 300 years ago (the motto of the Royal Society was 'Nullius in verba'). Nevertheless, until the 20th century there was generally little requirement on authors to justify their claims prior to publication, with the burden of proof generally being on opponents rather than proponents of ideas [79]. Benos et al. [20], citing Kronick [170], note that the first scientific journal, the *Journal des sçavans*, considered its role to be simply to report others' claims and findings rather than guarantee their accuracy. The early-20th-century *Annalen der Physik*, under Max Planck's editorial stewardship, generally allowed authors a great deal of leeway after their first publication [79], and this criterion ('published here before') still carries weight in the selection process of many journals.

Nevertheless, formal peer review as we understand it today still dates back to at least the 18th century. The *Philosophical Transactions* of the Royal Society of London—founded in 1665, the same year as the *Journal des sçavans*—was selective in its choice of manuscripts, but this was an informal process in the hands of the editor [269]. The Royal Society of Edinburgh's *Medical Essays and Observations*, first published in 1731, was probably the first to introduce peer review as we would recognise it today, with submitted manuscripts being distributed by the

editor to appropriate specialists for assessment [269, 20]; the *Philosophical Transactions* of the London society adopted this system in 1752 [269]. Different forms of review were adopted by other journals over the next two centuries, with some following this procedure of reports from recognised outside experts while others employed internal review panels. A few held out for a long time: the *Lancet*, one of the world's oldest and most highly-regarded medical journals, did not employ peer review until 1976 [20].

The present day ubiquitousness of peer review reflects this chequered past, with different journals[2] employing quite different practices of selecting and evaluating submitted articles. Broadly speaking, these tend to involve a mixture of editorial and reviewer-based selection. Journal editors are responsible for the primary decision of whether or not to submit a manuscript to review (some have a high rate of summary rejection) and, in the former case, to choose the most appropriate independent experts. They also have the final say in whether or not to accept referee recommendations—though it occurs only rarely, editors do sometimes publish against the advice of reviewers[3].

The main task of reviewers—selected researchers with (hopefully) some level of expertise appropriate to the claims and techniques of the submitted manuscript—is usually to ensure the technical correctness and clarity of the work, identifying methodological or empirical flaws and making recommendations for improvement where possible. More controversially, they are frequently asked to make judgements of significance and suitability for the journal: effectively, to make editorial decisions [177]. Many journals request recommendations for action[4], typically along the lines of 'accept', 'revise and accept', 'revise and resubmit' or 'reject', and these recommendations tend to be the key deciding force behind the final editorial decision whether or not to publish [61]. Increasing reliance on these referee judgements may provide one reason why many journals now request or require three different referee reports: the need to avoid a split decision [177].

Different journals' review procedures place different weight on these different aspects of the system, with some requiring only technical correctness[5] and topical suitability, whereas others place great emphasis on quality, innovation and significance. Some journals employ significant editorial selection, with a large proportion of manuscripts being rejected even before peer review. Some others may still editorially rule in favour of authors they know or respect without sending out for review, or at least treat them more kindly than less well known authors [146]. Besides asking for commentary and recommendations, some journals request authors to rate papers according to several criteria (quality, importance, etc.), for example on a 5-star scale. Lastly, journals differ in their policy towards review blindness or openness: most grant anonymity to reviewers, but a few prefer reviewers to openly sign their reports while others attempt to provide a 'double blind' procedure where neither authors nor reviewers are aware of each other's identity. This issue of

[2] Peer review is also employed in the assessment of funding applications and the evaluation of individual researchers and job applicants, and these different aspects of professional scientific life cannot entirely be separated: success in obtaining funding or employment often rests on prior publication records.

[3] Such articles are sometimes published as 'commentary' or invited papers or otherwise explicitly identified as not having passed through an independent peer review process.

[4] Not all journals prefer such concrete advice. Nature's peer review policy notes these options but adds that 'The most useful reports ... provide the editors with the information on which a decision should be based. Setting out the arguments for and against publication is often more helpful to the editors than a direct recommendation one way or the other.'

[5] Most fields have at least one journal where even this appears to be optional.

openness is one of the key foci of debate, ethics and research into peer review [112].

3.1.2 Shepherding

Shepherding[6] is a modification of the traditional peer-review process where a shepherd (reviewer) works together with the authors (sheep) of a paper to improve the paper. This process is non-anonymous. Shepherds are usually chosen among experienced authors participated in previous conferences and their task is to discuss the submission with sheep so that they can refine the paper prior to the conference. Shepherding usually results in several rounds of providing feedback and improving the paper. This process still includes the accept/reject decisions being made right after submission and after shepherding. Post shepherding papers may be accepted directly into a conference workshop (writers' workshop), or into a writing group. Writing group papers receive additional face-to-face shepherding at the conference itself, and those papers that reach the required standard are considered for workshop review on the final day of the conference.

Writers' workshops follow a special format which has been adopted from reviewing poetry. Before the conference, everybody reads each other's papers. In the actual workshop, authors give each other feedback on their work in a peer review fashion. Each writers' workshop contains 5 to 8 papers; a session of around an hour is devoted to each paper. In such a session, the authors of the paper under discussion remain silent while the other authors have a discussion about it, and explain what additional insights and views they have. Authors (as well as non-authors who may join) stay with their workshop over the entire conference. This way authors get a lot of ideas on how they can improve their work. Authors incorporate the feedback they receive at the writers' workshop into their papers before the papers go into the final proceedings after the conference.

It is worth saying that in line with the spirit of shepherding, the *PLoP conferences are structured differently from conventional conferences. Apart from writing groups and writers' workshops there are also focus groups, which are free-format discussion groups which bring together people who are interested in a challenging topics, and Birds of a Feather (BOF) sessions, which are spontaneous events, organized on site. Contents and format of a BOF session is up to the group joining the session. In addition to these, there are also game breaks that feature non-competitive games in order to let participants know each other and activate the creative halves of their brains, and build up a community of trust. Finally, all participants stay in the same, usually remote, place and have meals altogether.

Shepherding is applied in *PLoP (Pattern Languages of Programs) series of conferences, such as PLOP, EuroPLoP, ChiliPLoP, Mensore PLoP, KoalaPLoP, VikingPLoP, and SugarLoafPLoP. Papers discussed at writer's workshop at this conference qualify for submission to the journal "TPLoP - Transactions on Pattern Languages of Programming". This journal will be published by Springer.

3.1.3 Costs of peer review

The costs of publication need to be viewed in the context of the total costs of research communication. Different components dwarf the actual cost of publishing: first of all, the research itself, paid

[6]The text is partially adapted from `http://www.hillside.net/europlop/about.html`

by research funders, who, in a growing percentage of cases, are private organizations rather than public institutions or governments. Then, there is the time the researcher spends writing the paper, and this time is usually covered by her/his employers. Furthermore, the process of publishing the paper, once written, also has costs—both to the academic community (editing, peer review) and to the publisher. Given that, in our costs analysis we should distinguish between indirect costs of reviewing, usually covered by funders, which comprise the costs related to the resources spent in research activity and in writing the articles, and direct costs borne by publishers, which might concern editorial and coordination costs related to the reviewing process (highlighted in the last point).

Therefore, costs in research correspond to many different sub-activities, and also are consistently higher or lower depending on the people and the efforts involved. The quality check of research performed through reviews is sponsored either directly by funders or by collecting subscriptions' fees related to the publishing journal. This money, is often used to cover the costs of the journal: on the other hand, reviewers usually don't get any payment for their work due to different motivations. For instance, some of the large commercial journal publishers could afford to pay significant sums to individuals for their refereeing, but many journals in the humanities and social sciences are small scale, with print runs in the hundreds rather than the thousands. Such journals could not exist if they had to pay their referees. Similar points can be made about the difficulty of journals in the humanities and social sciences moving to an 'open access' model, in which contributors paid to have their papers evaluated for publication and referees were paid for their time. Under this model, journals would have to charge submissions at a rate to cover their costs. For the natural sciences and medicine, in which the costs of research are paid for out of a grant, it is possible to imagine something like a full cost charging regime. By contrast, the small scale, often unfunded, research in the humanities and social sciences could not survive under such a model.

Page et al. [221], estimate that the amount paid to editors (honoraria plus support costs) consists of 3-5 percent of the subscription income of a journal. Using their figures, a journal publishing 600 pages/year, say 60 papers, would have an income of about USD 300.000, so 3-5 percent would represent USD 9.000-15.000, or a cost of US 150-250 dollars per published paper. Referees are generally unpaid, though editors—and sometimes other members of editorial boards—receive an honorarium. We acknowledge here, that this analysis is somehow simplified. In fact, there is quite big difference between journals (even in the same discipline). Usually the more respected the journal, the higher the royalties of the editors (which are most often not the reviewers). Also, involvement of different associations usually bring up the costs, since they also want their share of the revenues.

Donovan [74] reported that one major scientific society employs a staff of about 25 individuals and spends about USD 3.5 million to process approximately 9.000 papers a year, which would amount to USD 400 per paper if all were acceptable. Since the rejection rate is 50 percent on average, the cost doubles to USD 800 for each publishable manuscript. From the small sample examined, Donovan concluded that peer review is expensive, with the cost for each manuscript submitted ranging between USD 100 and 400, *and for each paper published, between USD 200 and 800*.

Tenopir and King [285] found that the time spent (peer) reviewing article manuscripts was significant. Citing a variety of sources, they suggested that scientists were spending and average

of 6 hours reviewing rejected manuscripts and approximately the same amount of time reviewing successful ones. They also noted other studies that reported ranges of 3 to 5.4 hours.Based on their costing of researcher time, they suggested that peer review was costing around USD 480 per article. Citing Tenopir and King, Morris [205] suggested that peer review activities cost the academic community USD 480 per article in 1997—based on an average of 3-6 hours spent reviewing per article, by 2 or 3 referees—or around USD 540 at 2004 prices.

Another source of useful information on costs is Holmes' [140] article. Aldyth Holmes is Director of the NRC Research Press, the major Canadian publisher of scholarly journals. She maintained that the refereeing element of costs would be unaffected by the medium of publication (electronic, print or both) and quoted a 1996 figure (in Canadian dollars) of USD 41.80 per published page in *editorial office costs*. These include the editors' honoraria and the costs of editorial assistants based at the editors' institution, but no overheads. This amount represented about one-quarter of the total direct costs per published page (USD 169.93), to which was then added an approximately 100 percent level of overhead to make a final cost per published page of USD 331.49. As her figures are per published page they must include costs associated with rejected papers, so the base figure of USD 41.80 needs to be approximately halved to arrive at a figure per submitted page of say CAD 21. This would then be in fair agreement with Tenopir and King's [285] figure of USD 20.

3.1.4 Study and analysis

It is widely recognized that, for a subject of such crucial importance in professional research, peer review is far too poorly understood or studied [146, 159, 238, 267]. The views of many people might parallel those of the Nature editorial team that, as Churchill said of democracy, it is 'the worst system ... except for all the others that have been tried' [81]. The studies that have been carried out are varied, sometimes contradictory and, overall, extremely equivocal about peer review's effectiveness.

Research on review processes can, broadly speaking, be classified into a few general topics: effects on quality (does it achieve its commonly-stated aims of enhancing the quality of published articles and rejecting flawed or incorrect work?); biases and inconsistencies in its results; effects on innovation (does it inhibit the publication of novel ideas?); and differences between different review techniques and processes (open versus blind, trained versus untrained reviewers, and others). We give an overview of all these areas of investigation.

Quality enhancement and control

Two of the most commonly-stated purposes of peer review are to improve the quality of submitted manuscripts, and to identify error or deception [267]. Indeed, editors will often encourage reviewers to offer helpful or thoughtful comments even when it is clear the manuscript will be rejected [243]. Although both goals could be questioned[7], most of us would certainly like to read good papers that are technically correct, and assessing peer review's effectiveness in this respect

[7]For example, a technically incorrect but creative and inspirational paper may be of greater value than a technically correct but trivial or uninteresting one. Perhaps the journals for whom technical correctness appears optional are actually on to something.

should therefore be a key focus of research.

Concerning quality control, some notable incidents of fraud and incompetence contributed to the growth of criticisms in recent years towards the peer review system. For example, in September 2002, when Jan Hendrik Schon, tipped to be a Nobel Prize winner, was discovered to have published a series of fraudulent papers. Subsequently, 16 papers he had published were withdrawn from Nature, Science, Physical Review and Applied Physics Letters [180]. Similarly fabrication has been found in the life sciences: the German molecular biologists, Fridhelm Herrmann and Marion Brach, were accused of inventing data in forty-seven papers published in a number of prestigious periodicals [132]. Unfortunately, fraud is not the only issue: a study published on the effects of 'ecstasy' had to be retracted by Dr. George Ricaurte of Johns Hopkins University when it was realized that a more potent drug had been tested by mistake [239].

Goodman et al. [118] carried out a study on research articles accepted for publication in the Annals of Internal Medicine between March 1992 and March 1993, using a 34-item assessment instrument to measure quality[8]. Manuscripts were assessed before and after revision in response to peer review and the editorial process. General conclusions were that such revision resulted in modest improvement of the work, notably in terms of description and interpretation of the results (limitations and generalisability of the study and the tone of conclusions) and in reporting of confidence intervals for statistics. Poor manuscripts were improved to a greater degree than those which were already good when first submitted. However, the reliability of the assessment instrument was low and the results may be skewed by the fact that only manuscripts that were eventually accepted were studied[9].

A brief study by Purcell et al. [229] proposed a more general taxonomy for judging changes to manuscripts, with broad categories of flaws including 'too much information', 'too little information', 'inaccurate information', 'misplaced information' and 'structural problems'. Each category, except the last, had two or three sub-categories. Typical changes as a result of review were either the addition of missing information or the removal of extraneous material.

The broad implications of these studies are probably that peer review can provide assistance in matters of clarity, description and interpretation—ensuring that studies are well-described and that adequate information is provided for readers to interpret and understand them—as well as ensuring the reasonableness of inferred conclusions. In particular, these improvements rely on the specificity of reviewer commentary rather than global assessments [118]. On the other hand, the function of peer review as a mechanism of error detection is much more equivocal. Godlee et al. [113] and Schroter et al. [251] carried out different studies where deliberate errors (mostly major, some minor) were introduced into manuscripts already accepted by the British Medical Journal (BMJ). Only a small proportion were identified by reviewers; Godlee et al. [113] report that 16 percent of reviewers failed to find any of the mistakes introduced, and 33 percent recommended acceptance despite the introduced weaknesses, while the mean number of major errors

[8]Items included clarity of various issues (e.g. rationale, aims, study design), adequacy of procedures and precautions, appropriateness of methods, and others, all to be rated on a 1-5 point scale (with 'not applicable' an extra option).

[9]Several different interpretations could be placed on this point. It might be that the high selection criteria of the Annals of Internal Medicine (only 15 percent of submissions are published) mean that only manuscripts with very limited room for improvement are selected ('less selective journals may have the potential to improve research reporting even more'). On the other hand perhaps the peer review process remains able to offer only limited help to all manuscripts—the Annals employ a large editorial staff, so perhaps the level of assistance they can provide is greater than many other journals.

detected was 2 out of a total of 8. Schroter et al. [251] demonstrated that training could improve performance, but the overall rate of error detection remained low[10], with a mean 2 out of 9 major errors for untrained reviewers compared to 3 for those receiving training. Callaham et al. [47] reported a similarly low rate of error detection among 124 reviewers of the Annals of Emergency Medicine, who identified a mean 3.4 out of 10 major flaws in a fictitious manuscript[11].

One interesting but unstated result from all the studies would be whether reviewers collectively identified all the flaws in the various manuscripts. This could have strong implications for the potential effectiveness of an open community review process. For example, deliberate fraud generally proves difficult to identify in the conventional peer-review process—it is usually uncovered only after publication by the larger-scale attention of the wider scientific community [180, 102, 197, 89].

Bias and inconsistency

'The ideal reviewer,' notes Ingelfinger [146], 'should be totally objective, in other words, supernatural.' Scientists operate with limited time and knowledge and with (often strong) personal preferences about what good work consists of. They also work within institutional frameworks of hierarchy and status, and are part of a wider society which itself is far from free of prejudice. To what extent is peer review affected by these factors?

Early studies already reveal a number of interesting phenomena. Zuckerman and Merton [323], examining the archives of the Physical Review, found distinct differences in the treatment received by higher- and lower-status authors. Dividing the studied authors into three tiers[12], they noted that authors of the 'first rank' received typically much faster response times[13], a phenomenon possibly related to bias in the degree of editorial selection: only 13 percent of papers by first-rank authors were sent to outside referees, compared to 27 percent by those of the second rank and 42 percent for the rest. Acceptance rates for the three groups were, respectively, 90 percent, 86 percent and 73 percent. On the other hand, Zuckerman and Merton's study suggests that relative rank does not affect referee decisions—that is, referees return similar rates of acceptance whether they are 'outranked' by, outrank, or of similar status to the author whose work they are judging—and they note an inverse correlation between age and probability of acceptance, that is stronger for lower-status authors. Thus, their broad conclusion is that while prejudices surely play some part in the system,

[10] One possible reason for the low rate of error detection, suggested by Schroter et al. [251], is that reviewers give up looking for further errors after having found enough to reject the manuscript. However, the result of Godlee et al. [113] demonstrating that 33 percent of reviewers recommended acceptance despite the artificially-introduced weaknesses (compared to 12 percent recommending acceptance with major revision and 30 percent recommending rejection) may offer a counter-argument.

[11] The study was more generally an investigation of reviewer quality, demonstrating moderate editorial ability to identify good reviewers. When faced with the fictitious manuscript, highly-rated reviewers performed significantly better than poorly-rated reviewers.

[12] The first tier (a total of 91 authors) contained scientists who, by 1956, had received at least one of the ten most respected awards in physics (including the Nobel Prize and membership of the Royal Society or National Academy of Sciences). The second were some 583 physicists who had not won any of the aforementioned prizes but who had been included in the American Institute of Physics' archives of contemporary physicists. The third rank consisted of the remaining 8,864 authors.

[13] Forty-two percent in less than 2 months, compared to 35 percent and 29 percent for second- and third-tier authors respectively. Only 11 percent of first-rank authors had to wait more than 5 months for a response, compared to 20 percent and 30 percent for the other two groups.

the different treatment accorded authors primarily reflects differences in the quality of work.

Zuckerman and Merton themselves found little disagreement between referees, with only a small percentage clashing over the fundamental decision to accept or reject: two-thirds of the differences of opinion related to proposed revisions. On the other hand, they cite earlier studies by Orr and Kassab [219] on biomedical journals, and Smigel and Ross [265] on sociology, showing strong disagreement between referees occurring respectively some 25 percent and 28.5 percent of the time, compared to figures of 38 percent and 46 percent that would have been expected if decisions were made by chance.

These studies date, of course, from a time when peer review was employed much less frequently than today, when journals received far fewer submissions, and when science and scientists were arguably much less specialised and diverse than they are now: we cannot necessarily assume that their results still hold. However, they identify several key issues which carry over to analysis of the present day literature. First is the importance of distinguishing between bias that is the result of prejudice and bias that in fact results from underlying quality differences. Second is the degree of difference that can be observed, in practice and results, of the peer review process as employed by different disciplines and different journals. Indeed, even within a given field, studies of peer review may give contradictory results [154]. There may also be differences depending on what is being assessed, for example, research articles or funding applications [20].

Status or institutional bias has been investigated more recently by several authors. Benos et al. [20] cite a study by Ceci and Peters [54] suggesting that researchers from prominent institutions are favoured in peer review, and another by Garfunkel et al. [97] on submissions to the Journal of Pediatrics which suggested bias in the acceptance of brief reports but not regular research articles. A study by Link [187] indicates bias in favour of US-based researchers, strong where the referees themselves are US-based, weaker (but still present) when the referees are not themselves based in the US; however, the work does not take account of possible quality differences. Ray [234], noting the results of Link, also refers to a study by Nylenna et al. [214] where Scandinavian referees were sent versions of a short manuscript either in English or their own language: the latter was far more frequently rejected. Anecdotally, Ginsparg [103] notes that researchers from developing countries have credited the electronic arXiv preprint server with improving the consideration given to their research, feeling that previously they had suffered bias due to the low print or paper quality of their hard-copy preprints.

Gender bias is another frequent concern in peer review. Motivated by the disproportionate numbers of women leaving academic careers, Wennerås and Wold [311] investigated the peer review process of the Swedish Medical Research Council (MRC), one of the country's major funding organisations, for awarding postdoctoral fellowships. Their results revealed strong bias against female researchers, along with favouritism for researchers who were known to members of the MRC committee[14]. The two biases were of roughly equal magnitude, so that a female researcher known to a committee member might be judged at roughly the same level as a male researcher without such connection; but a female researcher not personally affiliated with any committee member would have to have a significantly greater body of high-impact work than a male colleague to gain an equal assessment.

[14]Swedish Medical Research Council policy prevents committee members from sitting in direct judgement on their colleagues or affiliates, but the supposedly neutral members tasked with this office seem nevertheless to be influenced, giving higher scores to researchers known to their committee peers.

Gender bias in the assessment of research articles appears more subtle in nature. A study by Lloyd [189] suggested that in fact, while male reviewers did not discriminate on grounds of gender, female reviewers strongly favoured female authors and perhaps were biased against male authors. However, this may be a result of the study being carried out in a female-dominated field, and Lloyd notes that the bias may be influenced by perceptions of authors violating sex-role stereotypes rather than (or as well as) gender per se. The apparent biases also become less significant if one considers only outright recommendations of rejection (as opposed to 'revise and resubmit') or acceptance (as opposed to 'accept pending revisions'), perhaps indicating that only initial review stages, rather than final acceptance rates, are subject to bias[15]. This latter contention is supported by the research of Gilbert et al. [100] examining back issues of *JAMA*, which shows different behaviours by male and female reviewers and editors but no final biases in terms of article acceptance. (The different behaviours themselves could conceivably be due to a more subtle form of bias—in the kind of articles assigned to male or female editors.)

Open versus blind review

The typical method of peer review employed by journals involves anonymous reviewers being asked to assess known authors. This practice has been called into question on a variety of grounds: lack of accountability, biases of various kinds, hidden conflicts of interest, and other forms of abuse or quality issues—as well as the basic ethical issue of whether it is fair for one party to the exchange to enjoy anonymity while the other is known. Depending on the particular concern, it may appear better either to disguise author identity—'double-blind' review—or to have an open review process where both authors and reviewers are known to each other. There also exist subtleties such as whether authors or reviewers should be allowed to make confidential comments to the editors, whether reviewer identity should be revealed during or only after the end of the review process, and so on.

Studies have focused on several aspects of these issues. To begin with, there is the question of whether these alternative techniques are in fact feasible. In particular, blinding reviewers to author identity[16] appears to be particularly difficult: McNutt et al. [198], in a trial carried out at the Journal of General Internal Medicine, discovered that in 27 percent of cases 'blinded' reviewers were able to identify the authors, while with Godlee et al. [113], studying articles submitted to BMJ, the figure was 23 percent (of 90 reviewers): in a more extensive study with 309 blinded reviewers, blinding was unsuccessful in 42 percent of cases [298]. Justice et al. [157], whose study covered several different journals, reported widely divergent rates of blinding success, but this might be due to the relatively small numbers of articles from each individual journal.

Cho et al. [59], in a follow-up study to Justice et al., examined reasons for unmasking, specifically concentrating on reviewer characteristics and aspects of journal policy. The only factor reliably predicting ability to identify authors was reviewers' individual research experience, including the number of years of review experience, number of articles published in recent years and percentage of time devoted to research. Katz et al. [161] examined articles themselves for

[15] Compare, for example, the differential initial treatment accorded authors of higher status, as reported by Zuckerman and Merton [323].

[16] There appear to be no studies on incidences of authors identifying anonymous referees, although there exist anecdotal accounts. In at least some cases these may be paranoia: editors at Nature have commented how, often, recipients of negative reviews will incorrectly assume the paper was blocked by a rival abusing the review process [67].

features that could indicate author identity, finding that out of 880 manuscripts submitted to two radiology journals, some 300 contained information that could potentially indicate to reviewers the identity of either authors, their institutions, or both. Editors of the journals, presented with the anonymised manuscripts, correctly identified the authors or institutions of 74 percent of the 300 potentially-unblindable manuscripts (corresponding to 25 percent of the total number of articles). The giveaway traits included, in decreasing order of frequency, authors' initials stated in the manuscript body (106 occurrences in the 300 articles), references to the authors' work in press (66), references identified as the authors' previous work (57), institutional identity present in one or more figures (54), institutional identity stated in the manuscript body (47), authors' names stated in the manuscript body (7), and authors' identity being revealed by either previously-published figures (7) or acknowledgements (4). At least some of these factors were violations (whether deliberate or accidental) of the journals' explicit instructions for manuscript submission, suggesting that journal policy is difficult to enforce in practice.

Open review, on the other hand, is easy (in principle) to implement, but carries with it fears of potential biases (e.g. towards high-status individuals or institutions) and the risk of increasing conflict or antagonism between authors and perhaps decreasing reviewers' willingness to be properly critical [112]. In practical terms, the main obstacle appears to be the unwillingness of some reviewers to participate. Referees in the study by McNutt et al. [198] were asked, but not required, to sign their names: only 43 percent complied. Van Rooyen et al. [297] note a difference (23 percent compared to 35 percent) in the proportion of participants declining to review, depending on whether anonymity was offered or not: of the latter, several gave as their explicit reason a personal opposition to open peer review. Walsh et al. [304] found that only 76 percent of referees approached were willing to sign their reviews.

Practical issues aside, the principal question all studies have addressed is that of quality—whether, and how, open or double-blind review procedures affect the review process. McNutt et al. [198] reported an improvement (in the opinion of editors) as a result of blinding referees to author identity, and no quality differences (again, according to editors' opinions) between reviews that were signed or unsigned by referees. Justice et al. [157] reported no quality difference (as perceived by editors and authors) between the reports of blinded or unblinded reviewers, but noted that this could be an effect of the lack of blinding success. The extensive study by van Rooyen et al. [298], using the validated Review Quality Instrument (RQI) [296] as a measure, reported no significant differences whether or not reviewers were blinded to author identity, and whether or not reviewer identity was hidden from authors[17]; nor were there apparent differences in the recommendations made. The same authors' subsequent study of open peer review [297], using the RQI, again found no statistically significant differences in quality or final recommendation.

Walsh et al. [304] conducted an interesting study which compared the quality and other factors of both signed and unsigned reviews of referees who had agreed to take part in an open-review trial (that is, agreed in principle to sign their reviews, while being asked to do so on a random basis) to those of referees who had refused to participate. Using the RQI as a measure, signed reviews were slightly higher in quality than unsigned ones, and considerably better than those given by referees who had refused to participate in the open peer review trial. Reviewers signing their reports were significantly less likely (18 percent compared to 33 percent) to recommend rejection. In contrast

[17] This result takes into account the fact that some 42 percent of 'blinded' referees were at least partially successful in identifying authors or institutions: the performance of truly blinded referees was similar to those for whom blinding was unsuccessful.

to van Rooyen et al. [297], it was found that reviewers signing their reports put more time into their review. Signed reviews were also significantly more courteous and less abusive in nature, although the majority of all reports were polite.

3.2 Research performance metrics and quality assessment

With thousands of scientists working across a huge range of ever-more-diverse disciplines, analytical methods and metrics to identify productive or important researchers are highly desirable. Detailed individual assessment being impossible on such a scale, funding agencies, research assessment panels and so on must of necessity rely either on personal contacts—carrying a strong risk of bias, nepotism and other negative influences—or attempt to identify proxy indicators of quality which allow quick and simple comparison.

On the other hand, such measurement and assessment techniques have come in for considerable and sustained criticism [177, 176, 178, 291, 48, 57, 287], primarily on the grounds that they are responsible for some very undesirable and destructive practices within professional science. Where a metric exists, scientists may be encouraged to follow research and publication practices that maximise their scores rather than providing the best service to the scientific community. Examples include an obsession with publishing in journals with a high impact factor, the division of research findings into 'least publishable units' so as to maximise the number of articles produced from a given project, and a whole range of political shenanigans ranging from citation swapping to abuse of the referee process.

It follows that when considering a potential research assessment metric, several questions must be asked. To begin with, does the metric reliably correspond to scientists' actual perception of quality [128][18]? Just as peer review may in some cases block rather than support the most interesting or innovative papers, so too metrics may overlook important research or researchers and promote uninteresting work. Second, is the metric open to abuse or manipulation? If so, then it is likely to encourage cheating and (often unethical) professional practices which damage and distort the scientific literature, favouring those who are willing to 'play the game' over those whose primary focus is doing good science [177, 176, 178, 89, 57, 287]. Other factors include whether the metric is self-distorting—that is, whether having a high score makes it easier to gain one still higher—and whether it carries implicit (or explicit) bias towards certain sections of the research community.

3.2.1 Citation Analysis

Two of the simplest and longest-standing metrics of scientific performance have been numbers of papers produced and the numbers of citations received.

An essential part of research papers, particularly in the sciences, is the list of references pointing to prior publications. As Ziman [321] observes, 'a scientific paper does not stand alone; it is embedded in the "literature" of the subject". A reference is the acknowledgment that one docu-

[18] An important consideration here is whether any one-dimensional metric can actually capture all the various facets of *quality* that are important to science. For example, most psychometric tests address multiple different factors, including both general and domain-specific measures [128].

ment gives to another; a citation is the acknowledgment that one document receives from another. In general, a citation implies a relationship between a part or the whole of the cited document and a part or the whole of the citing document. Citation analysis is that area of bibliometrics which deals with the study of these relationships.

The motivation of the usage of citation for measuring the relevance of publications is essentially that references could be considered as credits given to others for having somehow inspired further enhancements or new discoveries. Moreover, since citations are numeric values, it is easier to build, starting from them, more advanced algorithms to quantify how good or bad a certain work is.

The reasons these two metrics are valued is relatively clear: good scientists are likely to produce high volumes of work, and their work is more likely to be useful to (and therefore referenced by) their fellows. Even without any actual measurement or statistics, it is clear that a scientist not producing papers is professionally 'dead'—as is an article that is no longer being cited. From these raw numbers, a great variety of further measures can be obtained, and the use and interpretation of such measures is a question of considerable subtlety (see e.g. [30, 308, 294, 295, 179, 128]). For the purposes of this section, we will focus in particular on two that have gained great attention and popularity: the Journal Impact Factor [95, 93, 94] and the h-index [136, 137].

The Journal Impact Factor

Large-scale scientific citation analysis began in the 1950s [92] with the introduction of what would become the Science Citation Index now operated by Thomson Scientific. Initially the motivation appears to have been less focused on scientometrics than on providing a means for scientists to be aware of citations made to any given paper, and so facilitate the process of discovering what criticisms or extensions have been made to a particular piece of work. However, the construction of such an index—with extensive records of the citation network of papers—provided fertile ground for more quantitative analyses, the most well-known (and controversial) of these being the Journal Impact Factor (JIF), first introduced in 1963 [95, 93, 94].

Citations can be counted, and one way of measuring the quality of a publication to assess whether it appears in a high-citation outlet. For example, it is possible to ask whether an article appears in a journal with a high 'impact factor'. The chain of inference involved here runs as follows. The more papers in the journal are cited, the more impact that journal has. The more impact a journal has, the more authors will want to publish in that journal. The more authors who want to publish in the journal, the more demanding will be the selection criteria applied in the refereeing process. The more demanding the selection criteria applied in the refereeing process, the better the average paper will be. The better the average paper in the journal, the more it will be cited. And so a virtuous circle is completed. The initial motivation for the development of the JIF was simply to select which journals should be included in the Science Citation Index, and in particular, to develop a metric that would not simply favour journals with a high publication count [94]. The idea is simple: the impact factor is given by the total number of citations received this year by research articles (including reviews) published in the previous 2 years, divided by the number of such articles published in this time—that is, the mean citation rate per article within a given time window, following the notation of [300]:

$$GF_y = \frac{C_y}{P_{y-1} + P_{y-2}} \tag{3.1}$$

From these simple beginnings, the JIF has over time become one of the major tools of research assessment, with individual scientists frequently being assessed on the basis of the impact factor of the journals where they publish rather than the content or actual impact of their work itself [177, 176, 308]. This particular use of the JIF has come in for considerable criticism, including from journals who themselves have a high JIF. To begin with, the impact factor is not representative of individual article impact: journals with a high JIF typically do so because of a tiny minority of very highly-cited papers [254, 78, 80, 48]. Secondly, the JIF varies considerably across fields, and the predominant factor in determining its value appears to be simply the average number of citations in reference lists [6]. Thirdly, the long-term impact of papers may take decades to become apparent: Stringer et al. [273] note that the time scale can, for papers published in some journals, be as long as 26 years, and suggest an alternative ranking measure which takes into account each journal's individual transient period.

Obviously, there are further key points in the widespread usage of the impact factor that are particularly challenging, especially due to the multidisciplinarity of research assessment.

For instance, social science and humanities journals typically publish fewer articles than journals in medicine and the natural sciences. In consequence, journal rankings are sensitive to small number problems and can move erratically and significantly from year to year, often reflecting citations in relation to only one article. Whilst it can be argued that over time journal rankings will remain stable, there is also movement in those rankings (and it is highly desirable for incentive purposes that there should be) and inferences from ranked journal to author quality are not easy to make.

Assessment of impact is also complicated by the problem of time scale. The Thomson Scientific measure of two years reflects practices in the natural sciences and medicine, where specific results are picked up quickly in the literature. However, in the humanities and social sciences important papers may require greater time to be understood and absorbed than is true of papers in medicine and the natural sciences. Particularly striking examples of this phenomenon are to be found in the biographies of Nobel Prize winners in Economics. It took thirty years for the work of John Nash in game theory to be appreciated. The same is true of William Vickrey on tailor-made auctions of public assets.

The h-index

Jorge Hirsch proposed the h-index as a criterion to quantify the scientific output of a single researcher. The h-index was put forward as a better alternative to other citation-based metrics that could be used to measure research achievement (for example, total number of citations or citations per paper). Hirsch's h-index depends on both the number of a scientist's publications and the impact of the papers on the scientist's peers. Hirsch defined the index as follows: 'A scientist has index h if h of his or her N_p papers have at least h citations each and the other $(N_p - h)$ papers have $\leq h$ citations each' [136]. The index thus measures the broad impact of a scientist's work, rather than just productivity (a scientist can easily produce many boring papers) or raw citation numbers (in the index' terms, a scientist must be consistently highly cited across their output), and

generally correlates well with peer rankings [295]. That means that the h index favours 'those authors who produce a series of influential papers rather than those authors who either produce many papers that are soon forgotten or produce a few that are uncharacteristically influential' [162].

The h-index clearly has some limits. For example, one cannot have a higher h-index than one has total publications, and its value must be taken in context of the amount of time one has spent in research as well as other factors: Hirsch himself notes that, 'although I argue that a high h is a reliable indicator of high accomplishment, the converse is not necessarily always true' [136]. Hirsch has also suggested that, rather than a measure of quality for past work, the h-index might rather be seen as an indicator of future performance [137]. Lehmann et al. [179] claim that in fact the mean number of citations per paper is superior in this respect, though Hirsch [137] obtained different results. The h-index is also potentially open to manipulation by groups of scientists consistently cross-citing each other's papers [295], and clearly varies according to discipline [136], with, for example, scientists in the biomedical sciences having significantly higher values than those in the physics community. Thus, working practices with respect to output, authorship and citation may have a strong influence on h, and in particular this may imply not just discipline-specific variation in h values but also strong gender discrimination [280].

Other indices have been recently conceived to overcome the shortcomings of the h-index. Bornmann et al. [31] summarized them in their article and pointed out the improvements that these new indices would bring in the citation analysis. G-index, m-index, a-index, ar-index, r-index and m-quotient are some of the new indices which has been recently proposed. Basically, they are modification of the original h-index, which attempt to improve it on its 'weak' points, namely by giving more weight to highly-cited papers, or by considering the 'age' of the publications (i.e. how many 'recent' papers have been published and how was their impact).

Even if many of the new proposals provide for part of the bad points that characterize the h-index, and although the proposed variants may be conceptualized differently than the h-index theoretically or mathematically, in their empirical application they may be highly correlated with the h-index and with each other. Thus, we could say that the h-index, though it was one of the first proposals devised about citation-based metrics, still is the reference model adopted.

Performance of groups through aggregated bibliometric indicators

The performance of individual scholars may be aggregated at the level of groups of various sizes. Research groups, departments, and entire universities and corporations may thus be evaluated in pretty much the same way as that employed to assess the performance of individuals. According to Garfield and Welljams-Dorof [96], the value of using citation-based institutional rankings as science-and-technology indicators is obvious: "... university administrators and corporate managers can compare their peers and competitors. Government and private funding sources can monitor the return on their science and technology investment. And policymakers can identify relative strengths and weaknesses in strategically important ST sectors."

Noyons, Moed, and Luwel [213] describe their comparative evaluation of a Belgian research institute in micro-electronics, in which indicators of its own performance and that of its peer institutions were derived both from counts of citations received by publications written by members of those institutions, and from structural maps created using co-citation and co-word techniques. Vinkler [301] summarizes the applicability of a range of metrics, varying in degree of sophis-

tication, for evaluating the performance of research teams, differentiating gross indicators (e.g., raw counts of citations received) from specific indicators (e.g., number of citations per paper or per researcher), distribution indicators (e.g., proportion of total citations received by all research teams being compared), and relative indicators such as Vinkler's *Relative Citation Rate* (RCR)—the number of citations received, divided by the sum of the impact factors of the journals where the cited papers were published. This last metric is an example of a measure that compares counts of observed citations with estimates of some expected citation score, and is similar to the categorical journal impact used by ISI in their macro journal studies.

Van Raan [294] suggests that ranking of research institutions by bibliometric methods is an improper tool for research performance evaluation, even at the level of large institutions, because the indicators used for ranking are often not advanced enough. This situation is part of the broader problem of the application of insufficiently developed bibliometric indicators used by persons who do not have clear competence and experience in the field of quantitative studies of science. In particular, Van Raan describes some of the possible technical problems that could affect the process. This comprises problems related to the attribution of publications—and with that, of the citations to these publications—to specific organizations such as institutes, university departments, and even on a high aggregation level to the main organization, for instance universities. Indeed, it could happen that the main affiliation related to a particular author has not been specified or is not accurate. This is often caused by variations in the name of the same university, or because departments and institutes are mentioned without proper indication of the university. Furthermore, groups or institutes of a national research organization (such as the French CNRS) are quite often mentioned instead of the university where the research actually takes place. On top of that, further problems arise when two or more universities are within one city. In some cases these problems are so large (e.g., Vrije Universiteit Brussel and the Universit Libre de Bruxelles, both are indexed as "Free University (of) Brussels") that it is virtually impossible to distinguish both universities on citation-index based address. Moreover, we should not forget that aggregated indicators inherit all the shortcomings that characterize the evaluation of individuals or papers.

Critiques of citation analysis

Critics have questioned both the assumptions and methods of many studies found in the citation analysis literature. The strongest advocates of citation analysis recognize its limitations and exercise care in its applications. Unfortunately, other investigators seem to be unaware of these limitations and misinterpret the results of their analyses. The use of citation analyses for evaluative purposes is the issue that has generated the most discussion. While Bayer and Folger [15] note that measures derived from citation counts have high face validity, Thorne [286] argues that citation counts have spurious validity because documents can be cited for reasons irrelevant to their merit.

An interesting article by Linda Smith [266] describes what are the potential errors in the research evaluation through citation analysis induced by the misuse of citations. The assumptions, that should motivate citations, pointed out by the article are:

- Citation of a document implies use of that document by the citing author: otherwise, certain documents are underrated because not all items used were cited, and other documents are overrated because not all items cited were used

- Citation of a document (author, journal, etc.) reflects its merit (quality, significance, impact)

- Citations are made to the best possible works: documents cited do not necessarily represent the most outstanding in a particular field. It may be that anything which enhances a researcher's visibility is likely to increase his citation rate, irrespective of the intrinsic quality of his work

- A cited document is strongly related in content to the citing document: if this is not the case, the evaluation based on citations could be biased according to the "nature" of the publication we want to rank. For instance, a citation to a book which covers many different topics could be not as meaningful as that given to a very focused article, because, usually, in this case the citing article's content is not so strongly related to the cited publication's one

- All citations are equal: self-citations and citations coming from "known" people (from the cited publication's author perspective), usually, are not as relevant as those coming from people that have no relationships with the cited publication's author

Probably, until more is understood about the reasons for citing, citation counts should be viewed as a rough indicator of quality. Small differences in citation counts should not be interpreted as significant, but large differences may be interpreted as reflections of differences in quality and impact. Results of citation counts should be compared with alternative quality indicators to look for correlations.

3.3 New and alternative directions in review and quality promotion

An interesting point raised by Spier [269] in his history of the peer review process is that in many cases its adoption was technology-driven[19]. Only in the 1890s, with the introduction of the typewriter and carbon paper, did it become easy to make multiple copies of a manuscript; the almost universal uptake of the process in the second half of the 20th century can probably be linked to the introduction, in 1958, of the Xerox photocopier. Email, the internet and electronic documents have since facilitated the process still further. Yet this last technological revolution has opened up entirely new possibilities: to not just make traditional peer review faster and easier, but to dramatically change the way in which research is disseminated and evaluated [125, 216]. Among the novel developments are the rise of electronic preprint ('e-print') servers, the open access publication movement, the possibility of community review and commentary, and collaborative creation along the lines of Wikipedia and the free/open-source software community.

[19] Some of these technology-driven review processes were distinctly unpleasant in nature. As Spier points out, some of the first large-scale 'peer review' was conducted after the introduction of printing, when it became possible for the first time to mass-produce and widely distribute documents: obviously, ran the line of thinking at the time, it was necessary for someone to ensure that what was distributed met some basic standards. Unfortunately the 'peers' doing the review tended to be political and/or religious authorities whose sanction on research deemed worthy of rejection was rather more harsh than denial of publication.

3.3.1 Preprints, Eprints, open access, and the review process

Eprints

In terms of research dissemination, some fields have been making use of alternatives to journal publication for a long time. Paul Ginsparg, creator of the arXiv online preprint service, points out [103] that his system is simply an electronic continuation of a high-energy physics tradition dating back to the 1970s, when it became standard for research groups to post printed copies of their latest research articles to large mailing lists at the same time as they were submitted to journals[20]. The community would therefore receive the latest results months in advance of their refereed publication. Ginsparg notes that the community 'learned to determine from the title and abstract (and occasionally the authors) whether we wish to read a paper, and to verify necessary results rather than rely on the alleged verification of overworked or otherwise careless referees' [103].

Electronic preprint (or 'e-print') servers such as arXiv have changed the situation in a number of ways, not only greatly speeding and facilitating dissemination but also permitting long-term archival of documents, while drastically cutting costs compared to hard-copy delivery and storage [103, 216, 149]. The impact on a number of fields—notably physics and maths[21]—has been dramatic, and despite fears about the lack of quality assurance[22], the general standard seems comparable to that of the refereed journal literature [149] and may even be of slightly higher quality due to authorial self-selection [171, 69]. The latter may provide part of the reason why papers posted on the arXiv are on average more highly (and faster) cited than those not, although the phenomenon is probably due to a mixture of reasons, notably early[23] (probably more important than open) access [69]. Since arXival presence also leads to reduced downloads of the corresponding article from publisher websites, a further explanation may be that arXiv provides ease of access—a single portal to literature that in officially-published form is broken up across many different archives.

As several authors have noted [103, 125, 126, 216, 48], the possibility of self-archiving electronic manuscripts allows for some significant changes in the function of peer review. In the traditional world of print publishing, limits on storage capacity have meant that the primary function of the review process has been to assist editors with the problem of deciding, from an excess of submissions, what work deserves to be distributed [146, 269, 252]. With e-prints, the storage and distribution problem is solved and the practice of peer review need no longer be an entry condition but rather an option which can be employed with multiple different purposes: giving a mark of professional approval (perhaps required by funding or assessment agencies), adding commentary, or giving a quality mark which can go up or down; it could also be used to hierarchically select for attention or priority, much as the journal system does now[24]. Notably, the process of review no

[20] Some of the larger research groups might spend USD 15-20,000 per year on this activity in material and personnel costs.

[21] Among other examples, Grisha Perelman published his proof of the Poincar conjecture in a series of preprints on the arXiv and never submitted it to a conventional journal. As he later commented, 'If anybody is interested in my way of solving the problem, it is all there—let them go and read about it.'

[22] Quality control on arXiv is limited to an initial 'in-the-club' selection procedure—first-time uploaders must be sponsored by an existing arXiv author—and some moderation, mostly to ensure that papers are listed in the most appropriate subject areas.

[23] On the potential citation benefits of early access, see Newman [211] on the first-mover advantage.

[24] Harnad [126] notes, for example, the substantial shift in physics to using arXiv as the source of current research,

longer has to stop with the publication of an article [192, 70, 168, 167], but can be extended into a long-term post-publication process of discourse and continuous assessment.

Open Access

In order to make scholarly information more accessible and affordable, a number of alternatives, made possible with the technology of the internet, have been proposed. Some of them fall within the definition of what is called Open Access. In 2002, the Budapest Open Access Initiative[25] defined open access as the "world-wide electronic distribution of the peer-reviewed journal literature, completely free and unrestricted access to it by all scientists, scholars, teachers, students, and other curious minds."

An important definition of Open Access publishing comes from a meeting of the biomedical community held on April 11, 2003 in Bethesda, Maryland, and is commonly referred to as the Bethesda Statement on Open Access Publishing. It is composed of two clauses, one concerning copyright and the other concerning archival copies and access. An Open Access Publication is one that meets the following two conditions:

- The author(s) and copyright holder(s) grant(s) to all users a free, irrevocable, worldwide, perpetual right of access to, and a license to copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship, as well as the right to make small numbers of printed copies for their personal use.

- A complete version of the work and all supplemental materials, including a copy of the permission as stated above, in a suitable standard electronic format is deposited immediately upon initial publication in at least one online repository that is supported by an academic institution, scholarly society, government agency, or other well-established organization that seeks to enable open access, unrestricted distribution, interoperability, and long-term archiving (for the biomedical sciences, PubMed Central is such a repository).

Recently, open access gained considerable momentum, principally in the forms of open access publishing supported by grants and donations, author charges or other kinds of cost recovery, and open access archives and repositories. There is an important distinction between open access publishing (i.e. open access to formally published work) and open access archives and repositories, which may contain both formally published work (e.g. e-prints) and works that may not previously have been formally published. These alternative possibilities have led to some innovative publishing and review practices in the Open Access community. Beyond the immediate shift in the economics of publishing—charging authors (or their institutions or funding agencies) for the once-off costs of the editorial and review process, and making articles available to all without restriction—several have chosen to take further advantage of electronic distribution to develop novel ways of assessing and selecting research.

without reference to peer review, while at the same time relying on publication in the peer-reviewed journal system to provide marks of professional achievement.

[25] Budapest Open Access Initiative. `http://www.soros.org/openaccess/`

Indeed, in the open access publishing model, the costs of peer review and the production of journals are met from donations and/or institutional support, or wholly or in part by charging authors a per article or per page fee for publication, submission or some combination of both. These fees will be paid by the authors' institutions and/or funders, with publication regarded as a part of the cost of research. Currently, relatively few open access journals are author pays, with many using donations, bequests, institutional support, priced add-ons or auxiliary services to support publication. These models are still evolving and it is still relatively early to judge their role and viability with respect to other emerging and established models [141].

PLoS ONE, for example, employs conventional peer review but does so only to assess the technical aspects of a piece of work, and not its significance, novelty or subject area [192]. Published papers are then open to ongoing comments and rating by readers of the journal website. Among the reasons given for this practice are a desire to avoid the 'also-ran' phenomenon (where a high-quality article is rejected from high-profile publication because it has been 'scooped' by an existing paper), the need to foster interdisciplinary links rather than splitting the literature into ever-smaller topical specialities, and a simple recognition that 'importance' or significance often only becomes clear some time after a paper's publication. Where editorial selection is desirable, this can be provided by specialist access portals, which have the potential to be considerably more flexible and diverse than the relatively fixed selection criteria of many journals, able to serve both long-term and transitory areas of research interest. An even more liberal review scheme is provided by Biology Direct, an open access journal which is pioneering a novel form of open review [168, 167]. Authors select their own reviewers from the editorial board (although board members may ask an external expert to provide a review on their behalf), and instead of the typical journal requirement of positive referee reports, Biology Direct's only acceptance criterion[26] is that three members of the editorial board are interested enough by the paper to provide or solicit reviews. It is the author's choice whether or not to revise or withdraw the paper in response to critical comments, or to challenge referees' claims, and when a paper is published, the complete author-referee correspondence is published along with it. Thus, Biology Direct provides a publication scheme which reflects in many ways the more liberal and open discourse associated with scientific meetings: authors are provided with greater leeway in terms of the ideas they can share, but do so in the knowledge that they will be accompanied by critical commentary and discussion.

Open Access papers' impact

Interesting articles have been published to support the Open Access way of knowledge dissemination and to compare it to the traditional one. A recent Institute for Scientific Information (ISI) study has reported that traditional journals and Open Access journals have similar citation impact factors [227]. The ISI's press release announced: "Of the 8,700 selected journals currently covered in Web of Science191 are Open Access journals... [A study on] whether Open Access journals perform differently from other journals in their respective fields [found] no discernible difference in terms of citation impact or frequency with which the journal is cited". It is certainly remarkable the fact that there were almost no impact differences between the 191 Open Access journals and the 8509 non-Open Access journals indexed by ISI at that time, but to get a realistic estimate of the effect of Open Access on impact, it is not enough to compare only the 2 percent of ISI journals

[26] There is also an 'alert' system whereby reviewers can flag papers they believe to be pseudo-scientific rather than genuine research articles, and editors can reject on these grounds.

that are Open Access journals with the 98 percent that are not.

Brody et al. [129] report further data about experiments made to highlight the advantages (or, the non-disadvantages) of an Open Access approach from the researcher's perspective. The way they use to test the impact advantage of Open Access was to compare the citation counts of individual openly accessible and not-openly accessible articles appearing in the same (not-openly accessible) journals. What further needs to be compared is the citation impact of the much higher percentage—perhaps as high as 20-40 percent according to Swan and Brown's [279] sample—of articles from the 98 percent non-Open Access journals that have been made openly accessible by their authors with the citation impact of articles from those very same journals and issues that have not been made openly accessible by their authors. In this case, an article is made openly accessible by self-archiving it, which means by depositing a digital document in a publicly accessible website, preferably an Open Access-compliant Eprint archive.

They found that Open Access dramatically increases the number of potential users of any given article by adding those users who would otherwise have been unable to access it because their institution could not afford the access tolls of the journal in which it appeared; therefore, it stands to reason that Open Access can only increase both usage and impact. The ratio of "reads" to "cites" will no doubt vary by field. For example, Kurtz [172] and co-workers report it as 17:1 and even 12:1 in astrophysics. Odlyzko [215] predicts analogous trends in mathematics.

Another interesting study by Antelman [9] looks at articles in four disciplines at varying stages of adoption of Open Access—philosophy, political science, electrical and electronic engineering and mathematics—to see whether they have a greater impact as measured by citations in the ISI Web of Science database when their authors make them freely available on the Internet. The finding is that, across all four disciplines, freely available articles do have a greater research impact. Shedding light on this category of open access reveals that scholars in diverse disciplines are adopting Open Access practices and being rewarded for it.

Hajjem et al. [123] have tested Open Access (OA) articles' impact and cross-disciplinary generality using 1.307.038 articles published across 12 years (1992-2003) in 10 disciplines (Biology, Psychology, Sociology, Health, Political Science, Economics, Education, Law, Business, Management) gathering citation data from the ISI database. The overall percentage of OA (relative to total OA + non-OA) articles varies from 5-16 percent (depending on discipline, year and country) and is slowly climbing annually. By comparing OA and non-OA articles in the same journal/year, they found that OA articles have consistently more citations, the advantage varying from 36 to 172 percent by discipline and year. Comparing articles within six citation ranges (0, 1, 2-3, 4-7, 8-15, 16+ citations), the annual percentage of OA articles is growing significantly faster than non-OA within every citation range and the effect is greater with the more highly cited articles.

Figure 3.1, taken from Brody et al. [130] shows further data, partly taken from [129] and [123], which summarize the differences between disciplines with respect to the influence of the dissemination method (open vs closed) on the citation impact.

Growth of Open Access archives

As results confirming the striking correlation between access and impact become more widely known, a change in the way authors make their papers available can be anticipated. As most journals are not Open Access, authors will have two options. Wherever a suitable Open Access

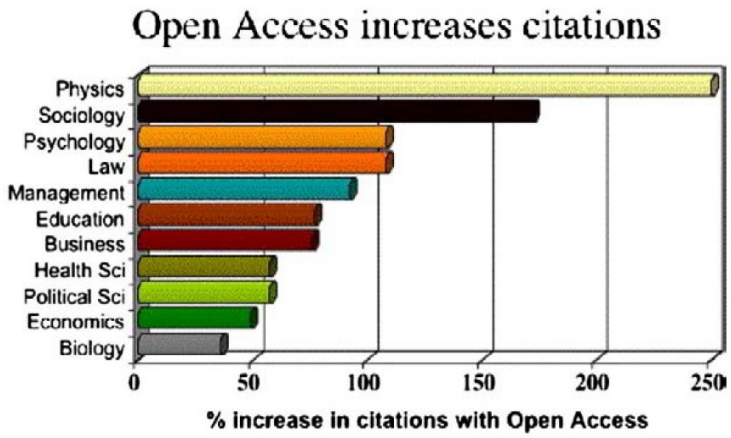

Figure 3.1: Average citation ratios for articles in the same journal and year that were and were not made OA by author self-archiving. Date span: 1992-2003. Sources: [123] and [129].

journal already exists for the subject matter of their article, authors can choose to publish in one of these. But, as stated by Brody et al. [138]: ”even according to the most optimistic estimates, less than 5 percent of the total number of refereed-journal articles published annually today (at least 2.5 million, in 24.000 journals) as yet have an open-access journal in which to publish them [138]”.

Most authors will continue to publish in established fee-access journals but they can in addition self-archive their papers in their own institution’s open-access e-print archives. An analysis of publisher-author agreements shows that almost 55 percent of journal titles from the publishers surveyed already ”explicitly left proprietary rights with the author” [91]. In other words, authors of papers in these journals can officially self-archive these papers. For the remaining papers not covered by such agreements, many of the journals will agree to self-archiving if asked. Concerning this scenario, Brody et al. claim in a recent article [130], that it is true that only about 10 percent of journals are Open Access, but also over 90 percent give their authors the possibility to self-archive; yet only about 10-20 percent of articles have been self-archived. To reach 100 percent OA, self-archiving needs to be mandated by researchers’ employers and funders, as they are now increasingly beginning to do.

3.3.2 Content evaluation in communities and social networks

During the last years we have witnessed a significant growth of the so-called internet communities, which essentially are groups of users that, communicating through the internet, share opinions, ideas, news, multimedia contents, etc. Starting from web forums up to social networks, almost all of those communities give their members the opportunity to evaluate the content published by the other registered users, and to provide a sort of feedback of relevance. To that extent, different type of techniques and metrics have been proposed, that could be taken into account if we are looking for new metrics to rank scientific and research-related content.

For instance, a possible alternative to citation count may be the number of times articles, put together in a web accessible library, have been bookmarked. This kind of operation is very popular in what is called ”social bookmarking”, namely the possibility given to every member of

the community to choose which objects they consider worthy of attention, by explicitly pointing them out. The aim is to create an online community in which members publish contents and share opinions about them with the other participants. The content shared could be of any kind, for example photos, web pages, news on the web, etc.

A typical feature of forums is the possibility to rank the contribution of a member by giving a mark to him/her, or to the content he/she has published. The mark is, sometimes, automatically computed, based, for example, on the number of discussions or messages posted, or assigned by other members depending on their opinion.

A similar kind of evaluation of members is the one introduced in Yahoo! Answers: here registered users make questions related to any topic they would like to know more about, in order, for instance, to solve problems or get the information they need. The promised reward for the best answer, obviously chosen by the applicant, is a number of points previously established: the greater the number of points collected by a member, the higher he or she is ranked.

eBay proposed another method of assessment, in this case about registered members. The system of evaluation introduced by eBay provides the chance to the buyer to give a feedback (positive, negative or neutral) to the seller. The feedback is useful for other users, that, looking at its value, can form an opinion about the reliability of the seller. So, it is a key feature that will increase or decrease the probability of selling our items, and, consequently, to make money. A further method often used as a criterion for ranking purpose is the number of times a web content has been downloaded. The act of downloading something is treated as an indication of interest, and, therefore, used as an index of relevance. Often, the number of times a page has been visited or the view count of a multimedia content (like a video) is taken into account. Indeed, many sites that offer video streaming (YouTube, for example) associate a counter like the one mentioned before to their content, in order to identify what are the preferences of the users (also for commercial purposes, i.e. web advertisement). Although download and view count are a bit misleading, because there is not a 1:1 matching between users and view count (namely, a single user might have visited a page multiple times), their usage is a very common practice in the internet.

3.3.3 Community review and collaborative creation

The attempts by PLoS ONE, Biology Direct and others to provide an alternative to conventional peer-review methods also reflect more fundamental changes that can be made to the way scientists share ideas and information. As Dayton [70] notes, while open access is important for all sorts of reasons, it still perpetuates significant inequality in scientific research: much key scientific knowledge and debate happens not on the pages of published articles but at scientific meetings or behind closed doors. Open discourse is an essential part of scientific advance, and whereas in the print publishing world the 'right to reply' tends to be in the hands of the few authors prestigious enough to get their commentary or letters published, electronic publishing makes possible continuous comment and debate on published material.

Such discourse-based, collaborative development processes have become the bread and butter of a number of non-scientific communities, with striking results. Most of these go far beyond commentary, feedback and ratings to allow full-scale community involvement in the creative process. Such practices are sustained as a result of several complementary factors including ethical philosophy, legal devices (in particular licensing) developed to support those ethics, and a variety

of technical tools that assist the collaborative process. Two such communities are worth examining in detail: the wide variety of projects coming under the umbrella of the free and open source (FOSS) software movements, and the two major collaborative encyclopedia projects, Wikipedia and Citizendium.

Free and open source software and distributed development

The Free Software movement was founded by Richard Stallman in the early 1980s as a reaction to the increasingly proprietary nature of software development—in particular, the practice of placing limits on the ways customers could use and modify software[27]. Stallman's response was to begin a project to create an operating system—the GNU system—which would grant the user exactly the desired freedom of use, but whose license would also constrain users to preserve those freedoms for others [272].

This licensing concept—later dubbed 'copyleft'—helped provide a legal framework via which diverse programmers, with diverse motivations and interests, could share and collaborate on software code. The power of these development practices were later highlighted by the Open Source movement for their strong practical benefits [235]: 'many eyeballs' to identify and fix bugs, and to propose and implement feature extensions.

The inevitable problem faced in such a collaborative environment is how to coordinate the diverse efforts of contributors. Even for single-developer projects it is important to be able to track changes to the code; where many developers are involved this becomes essential, with contributors needing to be able to keep track of the work of their peers, to review alterations to the source code and compare such changes to solved (or created) bugs in the program's performance.

The practices of different free/open source development communities vary considerably[28], but broadly speaking can be divided into two general development models, which reflect to a high degree the choice of tools to track the development of the software code. The most long-standing practice—reflected in version control systems such as CVS and Subversion—is of centralised development, where code revision history is stored on a single server to which a limited number of people have commit access[29]. Thus, all proposed changes to the code must be filtered and approved by at least one of these privileged individuals.

At the other extreme is the distributed model of development [260, 289], where developers operate essentially by peer-to-peer comparison and exchange of code. Pioneered particularly by the Linux kernel development team, this practice has becoming increasingly widespread as powerful distributed revision control (DRCS) tools—notably Bazaar, Git and Mercurial—have become available over the last few years. In contrast to centralised systems, these tools allow developers to create independent branches[30] (copies) of the revision history to which they can privately add: these changes can then be made available to others to merge (that is, incorporate into their own

[27] A parallel can be drawn to the ethical conventions of science, that researchers should share their results openly rather than keeping them secret for private gain: see for example `http://www.gnu.org/fry/`.

[28] See e.g. `http://bazaar-vcs.org/Workflows` for a discussion of some of the different development models.

[29] That is, they are able to make changes to the code.

[30] Centralised systems also allow branching, but branches can only be created in the central repository. Distributed revision control systems allow every user to create their own personal copies of the revision history on their own machine, more readily allowing independent development without disturbing the central repository.

copies) or ignore as they see fit.

In practice, most projects operate somewhere in the middle between these two extremes, and the particular advantage of DRCS is that it has made the precise nature of the development model (and thus the quality review system) a social choice rather than a technical requirement [289]. Centralised development has the advantage of allowing control over a project and its contents, but carries disadvantages of scale: as the number of potential contributors grows, so does the load on those with commit access, making it necessary to either restrict the development community size or widen commit access to the point where quality control may be weakened. DRCS, on the other hand, makes it possible for developers to operate highly independently, each having their own circle of well-regarded collaborators whose quality of work they have learned to trust, while saving their scrutiny for those whose work they do not know or have faith in. Changes to the code can therefore propagate via these 'circles of trust', reaching the 'trunk' branch controlled by core developers only after multiple rounds of scrutiny, revision and testing.

Wikipedia and Citizendium

The ethical philosophy behind the Free Software movement has spawned a wide variety of children: the Creative Commons and Free Culture movements, the Scientific Commons, and many textbooks published under free documentation licenses. One of the most well-known and successful is the community-created encyclopedia website, Wikipedia. Using the MediaWiki system for collaborative content creation, anyone may (anonymously, unless they wish otherwise) create or edit articles on any topic. Edits appear immediately in the published article and undergo no formal peer review.

Restrictions on participation are few—a small number of articles (for example, on contentious political topics) are protected to some degree, since otherwise they are too frequently vandalised, and while anyone can edit, only registered users can create new articles. Thus, in virtually all cases the only quality assurance is provided by the scrutiny of the unvetted contributing community. Despite this apparent lack of direction and control, in practice Wikipedia has been remarkably successful in generating a huge compendium of often very reliable information [101].

On the other hand, this same lack of direction and control means that whether or not material is accurate, it cannot be relied on as such [249], and consistent problems remain with bias, vandalism and lack of expertise. An alternative direction has been taken by the Citizendium project[31], which requires contributors to use real names and which employs a measure of expert review: while anyone can edit draft versions of articles, final versions require expert editorial approval. The approved version then remains the default presented to the public, but the latest draft is available to view if desired.

3.3.4 Recommender systems and information filtering

Thanks to the Internet and other computer networks, a large amount of customer opinions is now available both for academical and commercial use. As a result, nowadays we can see which movies have high average user ratings in a online movie database, web bookshops point us to new

[31] http://www.citizendium.org/

books according to our shopping history, Google uses our browsing and e-mail history to target advertisements, and so forth [250]. All these real-life examples are based, in one way or another, on recommender systems and information filtering in general.

When speaking about information filtering, web search engines [40, 165] (e.g. Google, Yahoo) present its landmark application in the age of the Internet. In its basic form, a web search engine provides a quality ranking of web pages and an efficient database that allows to quickly compare the search query entered by a user with the contents of locally stored webpages. An important drawback is that such search is not personalized: for a given search query, all users receive the same result. On the other hand, quality rankings of search engines can be successfully applied also in other areas—for example PageRank computed for the citation network of scientific literature can reveal influential papers [56].

When a record of past users' activities is available, personalized recommendation is likely to produce better results than 'recommendation for general audience' and this is the very aim of recommender systems. A recommender system is a specific type of information filtering that uses a limited number of user assessments of certain objects (books, movies, restaurants, etc.) to find which objects are likely to be appreciated by a given user. Apart from recommendation performance, among the issues that needs to be taken into account in a recommender system are data sparsity, large size of the data, noisy ratings, and spamming [135, 223].

At the heart of each recommender system there is a recommendation method which is used to process the input data. The first recommendation methods were popularity-based. In the case of explicit ratings (when users are asked to evaluate the objects in a given scale) this means that to predict the rating of user i for object α, either the average rating received by object α or the average rating given by user i can be used. While both approaches yield rather imprecise predictions, thanks to their low computational costs, the methods are widely used in practice. Moreover, the prediction by object-averages can be substantially improved if users' ratings are first aligned with each other by a simple linear transformation which makes the average rating and dispersion of ratings equal for all users [318]. Finally, in the case of implicit ratings (when users either include an object in their personal collections or not, no ratings are given), the analogue of object-averages is the assessment of object's popularity by the total number of users who has collected it.

A large number of recommendation methods are based on rating similarities between different users or different objects. That means, when recommending for a user, recommended are those objects that are liked by the users who rate similarly to the given user (we exploit user similarities) or recommended are those objects that are similarly rated as other objects already liked by the given user (we exploit object similarities). The latter approach was used in large-scale in the online shop Amazon.com [186]. In mathematical terms, denoting the similarity of users i and j as s_{ij} and the similarity of objects α and β as $s_{\alpha\beta}$, a similarity-based prediction of rating of user i for object α has the form

$$p_{i\alpha} \sim \sum_j s_{ij} v_{j\alpha} \tag{3.2}$$

when calculated according to user similarities, and

$$p_{i\alpha} \sim \sum_\beta s_{\alpha\beta} v_{i\beta} \tag{3.3}$$

according to object similarities.

When the number of users is much larger than the number of objects, object-based approach is computationally less expensive, and vice versa. While the basic idea is clear, much freedom is left in forming the exact equation used to obtain the rating predictions and, more importantly, in computing the similarities—for various choices see [259, 28, 281]. The standard way to decrease the computational complexity of the method and, in some cases, improve its performance, is to consider only k 'nearest neighbours' of a user (or an object) in the computation [115], such methods are known under the abbreviation kNN.

Another large class of recommender systems can be stamped as machine-learning techniques. These can involve content-based [222] or latent semantic [139] analysis, singular value decomposition [22], matrix factorization [281], and so forth (for an overview of machine-learning techniques see [3, 282]). In essence, they are all based on a plausible rating model with a vast number of parameters—their values are estimated by a multivariate optimization of the prediction error on training data (this is usually referred to as training procedure).

Finally, there are recommendation methods based on a transformation (projection) of the input data to a weighted object-object network and a diffusion-like process on the network. The idea behind the transformation is that whenever one user collects/rates two objects, there is probably some similarity between the objects and hence a link connected them is created or reinforced (see Fig. 3.2 and Fig. 3.3). The transformation is similar for both implicit and explicit ratings but in the latter case, the loss of information (which is always a side-effect for each projection) can be reduced if, instead of directly linking two objects, ratings given to these objects are linked. Consequently, recommendation for a particular user is obtained by propagating the opinions expressed by the user over the given network [318, 319, 320].

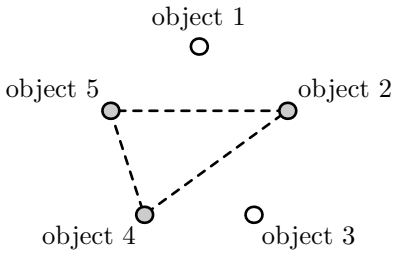

Figure 3.2: Implicit ratings: When there from the five available objects, a user has collected three of them, three links between the objects are created/reinforced.

When user's perception of an object is given mainly by object's quality and user's tastes play only a minor role, one can use the expressed opinions to deduce qualities of the objects. For example, the number of users who collected an object or, in the case of explicit ratings, the average rating given to an object, can be considered as crude measures of the object's quality. In practice, many users are good raters but some are misled, many users are honest but some are cheaters. To account with these influences, one can extend the system by assigning each user reputation modelled as a real-valued variable. Then qualities of all objects and reputations of all users can be estimated by an iterative procedure which lowers reputation of the users whose ratings diverge too much from the mainstream and when computing average ratings, gives high weight to users with high reputation [175, 71].

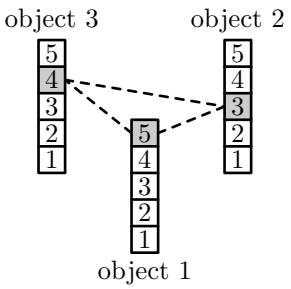

Figure 3.3: Explicit ratings: When a user has rated only objects 1 (rating 5), 2 (rating 3) and 3 (rating 4), three channels between the objects are created/reinforced.

3.4 Outlook

The present-day ubiquity of peer review as the gateway to publication has arguably been driven as much by the increasing volume of research articles as by any concern for quality assurance. The sheer scale of submissions received by some journals overwhelms the editorial team's ability to cope [146, 269, 252]. Given the widespread perception that peer review in practice creates a publication lottery [267], perhaps its use has more to do with the psychological need to have a system of at least apparent accountability: authors and editors need somebody to take responsibility for the decision to accept or reject a paper, to give reasons no matter how unfair or incorrect[32].

There are clearly strong positive aspects of peer review which we would like to maintain. Despite all the flaws identified, we should bear in mind the 'scores of scientists who have had their reputations saved by peer review' [83][33]. As Goodman et al. [118] point out, the principal value of peer review lies in specific comments and advice to assist authors in improving the paper rather than in general or abstract assessments of 'quality' or 'importance' on which there is usually little agreement. This coincides with the frequent experience of journal editors (and authors too) that, where performed well, peer review can offer a valuable source of collegial advice and support [238]. In revising the system of review and selection our aim should be to maintain possibilities for this kind of helpful assistance while attempting to offset the negative effects of the conventional review system such as delay, inconsistency and abuse.

The other major factor in the uptake of peer review has been a technological one [269]. Carbon paper, photocopiers, fax machines and electronic documents have all in turn simplified and speeded the process of sharing articles with reviewers. However, they have also been key factors in the increase in article production that is now such a problem. As a consequence it is not just individual editorial teams but the journal system as a whole that is overstretched and overwhelmed. The hierarchical system of journals, from highly-selective to highly-permissive and from general to specialist, has reached its own limits of capacity: the literature surely cannot sustain indefinitely the current process of fragmentation into ever-more-specialist titles.

Fortunately, the electronic technology that has led to this explosion in article and journal

[32]'Peer review represents a crucial democratization of the editorial process, incorporating and educating large numbers of the scientific community, and *lessening the impression that editorial decisions are arbitrary*' [our emphasis] [238]. See also the discussion in Ingelfinger [146].

[33]We should perhaps note that we do not have statistics, even anecdotal ones, on how many scientific careers have been unnecessarily or unfairly *destroyed* by peer review.

production also offers new means of production and distribution that in turn open up new (and possibly improved) methods of review. With the huge storage and distribution capacity of electronic archives it is now possible for diverse articles to share a common repository, as with arXiv and PLoS ONE, with topical relevance indicated by tags, keywords and good search and recommendation tools rather than crude and unbreakable divisions as with paper journals. While pre-publication peer review remains an option it need no longer always be a requirement: information filtering tools offer the opportunity to leverage the opinion of the wider community in judging a work, with top articles being identified not by their point of entry into the research literature—as tends to happen now with top journals—but by their impact and appreciation by the research community as a whole.

Information technology also offers new means of collaboration. Version control tools and collaborative creation systems such as Wikis open up the possibility of whole scientific communities working together on a common body of work. Instead of individual groups each publishing their own papers with incremental improvements, successive updates can be collectively reviewed and incorporated into the shared work. The process of research scrutiny and assessment would then be a community activity running hand-in-hand with the ongoing research process, just as it is already in the internal functions of existing research groups.

In summary, new electronic distribution and development tools offer a wide range of alternatives to the present system for both review and selection of research. Most of these methods are complementary, and together they open up the possibility of a much more collaborative, cooperative process of scientific research and assessment.

3.4.1 Further reading

The present chapter has covered a wide range of material in a relatively short space. Readers interested in a more in-depth view of some of the topics may like to peruse some of the following selected articles.

Peer review

Spier [269] and Benos et al. [20] provide good brief histories of the peer review process and (in the latter case) a review of the major issues and accusations surrounding it; the review by Ingelfinger [146] is now somewhat out of date but makes good reading to see how the situation has changed in the last 30-40 years. Dalton [67] and Lawrence [177] offer good descriptions of the 'social' problems and consequences of peer review.

Several special issues of JAMA have been dedicated to study and analysis of the review process: JAMA [150, 151, 152, 153]. Nature has also dedicated a Web Focus debate to the topic.

Research assessment metrics

The various articles by Garfield [92, 93, 94] provide an interesting history of the conventional journal impact factor that he played the major role in developing, while Lawrence [176, 178] offers a strong critique. Lehmann et al. [179] and van Raan [295] offer some comparisons of different

quality assessment methods. *Ethics in Science and Environmental Politics* and *Marine Ecology Progress Series* have published theme sections on, respectively, 'the use and misuse of bibliometric indices in evaluating scholarly performance' [42] and 'quality in science publishing' [41].

Information filtering

Hanani et al. [124] give a good overview of the field of information filtering, its challenges and concepts. Adomavicius and Tuzhilin [3] and Perugini et al. [223] provide good reviews of the recommender system research literature, while Brusilovsky et al. [44] contains several interesting articles on different aspects and types of recommender systems. Mizzaro [203] offers an interesting perspective on applying information filtering to quality assessment in the research literature.

Collaborative creation

Talks by Mark Shuttleworth [260] and Linus Torvalds [289] offer interesting perspectives on the tools and social practices of distributed collaborative development in the free and open source software communities. MacCallum [194] and Dayton [70] suggest ways in which scientific discourse can be improved through extensions to open-access publication.

Part 4

Computational trust, reputation models, and social network analysis

Being able to automatically provide a 'fair' credit attribution for both liquid publications (or SKOs) and researchers (such as authors, reviewers, etc.) alike is crucial to the success of this new publication paradigm.

Numerous trust and reputation mechanisms already exist in the literature. These usually rely on several sources of information for computing reputation, which may roughly be grouped into three categories [232]:

- Direct information from personal past experience with the peer in question
- Indirect information from the past experience of third party members of the network with the peer in question
- Socio-cognitive information, such as the belief in the willingness, capability, and persistence of the peer in question in carrying out a certain action.

To our knowledge, the majority of research in this field has focused mainly on the outcome of previous interactions, whether this information has been obtained directly from personal experience or indirectly from the experience of other members in the network. Only a few have focused on other sources of information. For example, Castelfranchi and Falcone [53] and Brainov and Sandholm [38] have focused on a cognitive view of trust.

In the LiquidPub paradigm, we believe it is critical to guarantee the fairness of the system by detecting redundant ratings, consistent biases, deliberately distorted ratings, unreliable reviewers, etc. To achieve this, we believe social relationships, along with other social network measurements, should be considered in addition to direct and indirect information from previous experiences.

Hence, the first section of this chapter provides a background on the available literature on computational trust and reputation models, which are mainly based on direct and indirect experiences. Since we believe social network analysis will be crucial for our work, the second section of this chapter provides an introduction to the available literature that makes use of social network analysis for measuring trust and reputation.

4.1 Computational Trust and Reputation Models

The scientific research in the area of computational trust and reputation mechanisms is a recent discipline oriented to increase the reliability and performance of electronic communities by introducing in such communities these well known human social control mechanisms. There are several reviews in the literature [247, 232] that analyse and compare the increasing amount of models that have appeared during the last few years. These models, apart from the mechanisms used to calculate the trust and reputation values, also differ in the very essence of what is trust and reputation. In this section we will present a general taxonomy that will help us to determine which are the type of models that are more interesting for the LiquidPub project. Finally we will present briefly some of these models.

There are two kinds of social evaluations that play an essential role in trust and reputation models: image and reputation. Both social evaluations concern other agents' (targets) attitudes toward socially desirable behaviour, and may be shared by a multitude of individuals. Image is an evaluative belief and it tells that the target is "good" when it displays a certain behaviour, and that it is "bad" in the opposite case. Reputation is instead a shared voice, i.e. a belief about others saying that a given target enjoys or suffers from a shared image. In other words, reputation is true when it is actually spread, not when it is accurate. From now on, we will define reputation as the group opinion on (what is said about) someone (or something) playing a specific role.

Image is subjective by nature. Two individuals, even if they have observed the same interactions can have a completely different image of a target. Image depends on the personal experiences of the individual but also on other internal elements like for example the goals the individual has. In the case of reputation we find two ways of looking at it:

- What we call subjective reputation, which as the name suggests, is also subjective in a similar way image is. Models that consider reputation as a subjective property assume that every individual can have a different method to calculate the reputation values. Also, you cannot assume that all members of the society have the same knowledge. Given that, the reputation value will depend on who is calculating that value. Examples of models that follow this approach are ReGreT [245], RepAge [248], Sierra-Debenham model [262], AFRAS [49], FIRE [144] among others.

- What we call global reputation. In this case, the approach assumes there is a shared (and agreed) method to calculate the reputation values and that this calculation is performed over the same set of elements that are public and therefore available to all the individuals. In this case, we can say that each individual has a public reputation in front of the society because the calculation does not depend on who is the evaluator. Usually, these models rely on a central service that is responsible for calculating the reputation values, although this centrality is not strictly necessary if we can guarantee that each individual is using the same method and data to make the calculations. Examples of models that follow this approach are those used in on-line auctions like eBay[1] or Amazon Auctions[2], laboratory models like Sporas [317] and web related methods (based on network analysis) like PageRank [7], HITS [164], or TrustRank [122].

[1] `http://www.ebay.com`
[2] `http://auctions.amazon.com`

In the LiquidPub project we are looking for mechanisms to rate different elements that go from SKOs to reviewers or authors. These ratings must be public and transparent so any individual knows where they come from so they can redo the calculations obtaining exactly the same results. Given that, what we need in the LiquidPub project is what we have defined as global reputation models. Both image and subjective reputation are not useful in this specific context.

What is clear also is that a single method to calculate reputation is not enough. Even for calculating the reputation of an individual in a specific role we will have to provide different mechanisms to calculate the reputation, each one stressing a different aspect of the interactions. This set of methods has to be public and available for every individual.

We will present now some the models that we think can be relevant in the LiquidPub project.

4.1.1 Online reputation mechanisms: eBay, Amazon Auctions and OnSale

eBay[3], Amazon Auctions[4] and OnSale Exchange[5] are good examples of online marketplaces that use reputation mechanisms. eBay is one of the world's largest online auction sites. Most items on eBay are sold through English auctions, where the auctioneer announces a reserve price and afterwards accepts increasingly higher bids. The bidder with the highest bid wins the item for the value of its bid. The reputation mechanism used is based on the ratings that users perform after the completion of a transaction. The user can give three possible values: positive(1), negative(-1) or neutral(0). The reputation value is computed as the sum of those ratings over the last six months. Similarly, Amazon Auctions and OnSale Exchange use also a mean (in this case of all ratings) to assign a reputation value.

All these models consider reputation as a global property and use a single value that is not dependent on the context. The information source used to build the reputation value is the information that comes from other agents that previously interacted with the target agent (witness information). They do not provide explicit mechanisms to deal with users that provide false information. A great number of opinions that "dilute" false or biased information is the only way to increase the reliability of the reputation value. Dellarocas [72] points out that the commercial success of online electronic markets suggests the models have achieved their primary objective: 'generate sufficient trust among buyers to persuade them to assume the risk of transacting with complete strangers'.

Certainly these reputation mechanisms have contributed to the success of e-markets like eBay but what is not clear is to which extend. There are several studies that try to analyse the properties of these models specially based on eBay data sets (see again [72]).

4.1.2 Sporas

Sporas [317] is an evolved version of the online reputation models. In this model, only the most recent rating between two users is considered. Another important characteristic is that users with very high reputation values experience much smaller rating changes after each update than users

[3] http://www.ebay.com
[4] http://auctions.amazon.com
[5] http://www.onsale.com

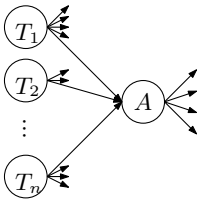

Figure 4.1: The nodes and links affecting the PageRank of node A

with a low reputation. Using a similar approach to the Glicko system [111], a computational method used to evaluate the player's relative strengths in pairwise games, Sporas incorporates a measure of the reliability of the users' reputation based on the standard deviation of reputation values. This model has the same general characteristics as the previously commented online reputation mechanisms. However, it is more robust to changes in the behaviour of a user and the reliability measure improves the usability of the reputation value.

4.1.3 PageRank

PageRank [7] is a patent by Google. The mechanism is inspired by how the number of citations determine the relevance of a paper in the scientific community. As the authors say in the description of the patent "PageRank is a method that assigns importance ranks to nodes in a linked database, such as any database of documents containing citations, the world wide web or any other hypermedia database". The main idea behind PageRank is to interpret a link from one page x to a page y as a vote from page x for page y. PageRank also takes into account the PageRank value of the page that casts the vote. As illustrated by Figure 4.1, the PageRank of a node A is affected by the PageRanks of the nodes linking to A, as well as the number of outgoing links each of those nodes has. Similarly, the PageRank of node A, along with the number of outgoing links from A, will in its turn affect the nodes that node A links to. The formula used by PageRank is the following:

$$PR(A) = (1-d) + d(\frac{PR(T_1)}{C(T_1)} + ... + \frac{PR(T_n)}{C(T_n)})$$

where:

- $PR(A)$ is the PageRank of page A,
- $PR(T_i)$ is the PageRank of pages T_i that link page A,
- $C(T_i)$ is the number of exit links from page T_i, and
- d, which is illustrated below, is a constant between 0 and 1.

PageRanks essentially models a *random surfer* who is randomly clicking on links, yet eventually stops. d is the damping factor representing the probability that this random surfer will continue clicking. We note that the mechanism is analogous to the steady state of random walkers on hyperlink networks.

4.1.4 HITS

Like PageRank, the HITS [164] mechanism is used to evaluate the relevance of a web page according to the link's network. However, instead of using all the WWW like PageRank, the link relations that are taken into account are those associated with what they call 'authority' pages. This idea is based on the 'Abundance Problem', that is, "the number of pages that could reasonably be returned as relevant is far too large for a human user to digest". Therefore, instead of using any page that links to the target page, we first select a subset of 'authority' pages and these subset is what will be used to make the calculations. The authors present an algorithm to determine this subset of pages. They define also the concept of 'hub'. A 'hub' is a page that have links to multiple relevant authoritative pages. Given this, each page has associated two values: its authority value and its hub value.

However, HITS has peculiarities. It is executed at query time, not at indexing time, with the associated performance cost adding to query-time processing. It is not commonly used by search engines. Also, HITS computes two scores per document, hub and authority, as opposed to a single score.

4.2 Social Network Analysis for Trust and Reputation

Social network analysis may be defined as a set of methods for analysing social structure by investigating relational aspects of these structures. Social networks are composed of vertices (or nodes) and edges. A vertex is the fundamental unit of a network. It usually represents people or groups of people (e.g. organizations). An edge is a link connecting two vertices. It represents relational data, such as kinship, friendship, colleagueship, scientific collaboration, etc.

Different types of networks may exist based on the different types of vertices and/or edges. For instance, vertices in a network may be either of the same type or different types (e.g. representing different nationalities). Similarly, edges in a network may also be of the same type or of different types (e.g. representing friendship and animosity). Additionally, vertices and edges may be given weights. For example, a weight of a vertex may represent how important is the person in its community. A weighted edge may provide better insight on the strength of the friendship between two people. Finally, edges may also be directed, i.e. pointing in one direction. This would be useful, for example, to represent the direction of email messaging between people.

In our research, we are interested in the analysis of such networks. For an overview of the most common network properties that may be computed, see Newman [210]. In what follows, we present an overview of the selected literature on social networks and reputation that we believe is most related to our work in the LiquidPub project. The selected mechanisms use various types of social network analysis for computing some sort of 'reputation' measures.

4.2.1 Finding the Best Connected Scientist via Social Network Analysis

Newman [212] focuses on the properties of coauthorship networks for choosing the best connected scientist. The study focused on the Physics E-print Archive database from 1992 to the present, the Medline database from 1961 to the present, the SPIRES database from 1974 to the present, and

the NCSTRL database for the last ten years.

The following basic results where automatically computed:

- number of authors
- number of papers per author
- number of authors per paper
- number of collaborators per author
- size of the giant component (i.e. the connected subset of vertices whose size scales extensively), and the percentage of the entire network that this giant component covers
- clustering coefficient of the network
- shortest path between two nodes
- betweenness centrality of a node
- average distances between nodes

The paper also shows how weighted collaboration networks may be used to measure the closeness of collaborative ties and who is the best connected scientist. The idea is that the number of papers each pair of scientists coauthored and how many other coauthors are on those papers affect the strength of coauthorship ties. The distance between two scientists becomes an inverse of the weight of their collaborative ties. This new distance definition is then used in computing the best connected scientist.

4.2.2 Searching for Relevant Publications via an Enhanced P2P Search

In the work done by Chirita et al. [58], a P2P search strategy is enhanced with social network analysis to improve the search. Their work was applied to scientific collaboration networks to improve keyword search for relevant publications.

For selecting the peers to forward the query to, three strategies are introduced. The first makes use of the peer's connectivity in the network. The second makes use of the peer's reputation. This could be a Pagerank or any other simple metric. The third makes use of the similarity between the chosen peer and the querying one. Hybrid models may then be constructed using these three models.

The results show that the best outcome is achieved when choosing peers based on the relative similarity ratio. Also, increasing the number of chosen similar peers to a number larger than 3 does not provide an impressive increase in performance. Finally, it turns out that combining the similarity and connectivity based strategies does not improve performance! This implies that the best connected peers are not necessarily the repository of the best resources in the network.

4.2.3 Using Social Relationships for Computing Reputation

In the work done by Sabater and Sierra [246], social network analysis, which is based on studying the social relationships between peers, is used to provide an additional source of information (to traditional information obtained from direct interactions and those obtained from other members of the society about their past experiences) for computing reputations. The basic idea of the system is that reputation is based on three dimensions: the individual dimension, the social dimension, and the ontological dimension. The individual dimension models the direct interactions between two agents. In this dimension, the subjective reputation is calculated directly from an agent's impressions database. However, the reputation value on its own is not sufficient. Hence a reliability measure of this value is also calculated by taking into account the number of impressions used to calculate the reputation value and the variability of their rating values.

The social dimension models the case when the source of information comes from other agents in the system. Depending on the information source, three types of social reputations arise: witness reputation, neighbourhood reputation, and system reputation.

In witness reputation, to identify witnesses, first the most connected sub-graphs are obtained, then the nodes with local centrality are identified. This way the peer will end up using only the most representative agents for its source of information; hence, it will be minimizing the correlated evidence problem. The next step would be to aggregate all this witness information. The final reputation measure would be an aggregation of the reputation values provided by all witnesses, taking into account how much trustworthy is each witness in providing its reputation value of the target agent. The final reliability measure is also an aggregation of the reliability and trust measures of each individual witness. The trust measure is defined as follows. The degree of trust that an agent a has in agent b on providing feedback about agent c is a combination of subjective trust reputations, which are calculated in a similar manner to individual reputation, and social trust. As for social trust measures, these are the results of applying fuzzy rules. These fuzzy rules depend on the type of relationships in a specific scenario, such as competitive, cooperative, and trade relationships. Note that different fuzzy rules would exist for different contexts and scenarios.

The basic idea of neighbourhood reputation is that the reputation of a peer's neighbour, along with the type of relationship that exists between the two (e.g. competitive versus cooperative relationships), can give an idea about to the reputation of the target peer itself. Again, fuzzy rules, which are domain dependent, are used to generate the reputation of the target peer based on the reputation of its neighbour and their relationship type. The reputations obtained from investigating the reputations of all neighbours are aggregated, using the reliability of each neighbour, to obtain one final reputation value. Similarly, the reliability of all neighbours are also aggregated into one final reliability value.

As for the system reputation, it is a default reputation value obtained from the observable role an agent plays in a given institutional structure. It is domain dependent and part of the initial knowledge of the agent. This observable feature is what differs system reputation from other social structures.

Finally, different reputation types may combine in such a way resulting in new reputations. For instance, to have the reputation of a swindler seller, one should have the reputation of overcharging items and the reputation of delivering items with a poor quality. This model is what the authors refer to as the ontological dimension. When one reputation concept is the parent node of two or

more reputation concepts in the ontological tree, then both the reputation and reliability measures of the parent node become a combination of the reputation and reliability measures of the children nodes.

The final reputation and reliability measures are defined as a combination of those of the individual dimension, the social dimension, and the ontological dimension. Agents have the option of deciding which dimension may be more relevant. For instance, if the agent does not have enough information, then it can decide that the social dimension will be most relevant.

4.2.4 Identifying, Categorizing, and Analysing Social Relations for Trust Evaluation

As we have seen above, more expressive reputation valuations may be achieved if the relationship between the agent in question and its recommending agent is considered. This is opposed to traditional techniques that assume that recommending agents are fully trusted. Ashri et al. [10] propose a method for dynamically identifying these relationships, categorizing them, and analysing them.

In this model, agents are defined as 'entities described by a set of attributes'. Attributes are features of the environment, and agents are capable of performing actions by adding or removing these attributes. Agents also pursue goals, which are achieved by performing actions. Agent actions are divided into sensor capabilities, which retrieve the values of the environment's attributes, and actuator capabilities, which change the environment's attributes. Attributes that may be manipulated by an agent are defined as the agent's region of influence (RoI). Those that may be sensed by an agent are defined as the agent's viewable environment (VE). As a result, goals are divided into query goals that require an agent to perform a sensory action that lies within the agent's VE, and achievement goals that require an agent to perform an actuating action that lies within the agent's RoI. According to this model, relationships between agents may then be identified by studying how their VEs and RoIs overlap.

Relationships may then be characterized into several types, depending on the context. The relationship types illustrated by Ashri et al. [10] lie in the context of an e-commerce scenario. First, two agents are said to be in a trade relationship if the goal of agent a is to sell a product that agent b may or may not wish to acquire. Second, one agent may be dependent on another, for instance, if agent a is selling a product that agent b wishes to buy. The intensity of the relationship depends on several factors, such as the number of sellers, the abundancy of the product, etc. This intensity basically determines who is dependent on whom and to what extent is it dependent on the other. In this case, the goal of the depending agent lies in the other agent's RoI; moreover, the other agent's RoI lies in the intersection of both agents' VEs. Third, two agents may be in a competitive relationship, for instance, if both are selling the same products. Again, the intensity of this relationship may depend on several factors, such as the price of the product, the market share, etc. In this case, either the agents share goals that lie in the intersection of their VEs, or they share RoIs that lie in the intersection of their VEs. Fourth, two agents may be in a collaborative relationship, for instance, if agent a is selling goods to agent b while b is also selling goods to a. This is a combination of two trade dependency relationships. In this case, the goals of b lie within the RoI of a, the goals of a lie within the RoI of b, and the RoIs of both a and b lie in the intersection of their VEs. Finally, relationships between more than two agents may be considered. The resulting configuration would be composed of the four configurations presented

above. However, some exceptions may arise giving special privilege to one agent over the other.

Computing trust and reputation may now be modified to accommodate the additional information obtained from identifying relationships. Sample rules are provided. For example, in the case of having a dependency relationship between two peers, the initial trust value is set to an arbitrary low value (in other words, confidence is low) if the intensity of the relationship is high, while it is set to a high value if the intensity is low.

4.2.5 Revyu and del.icio.us Tag Analysis for Finding the Most Trusted Peer

The main goal of Heath et al. [134] is to find out who knows what and who is the most trustworthy in delivering information on a specific subject. To achieve this, topic experience profiles are generated for each peer using 'Revyu', 'del.icio.us', and FOAF descriptions.

Previous empirical studies showed that people usually based their trust on the following five factors: expertise, experience, affinity, impartiality, and track record. The different factors were selected based on the criticality of the task and subjectivity of possible solutions. However, the first three were given much more emphasis than the others. Hence, the research of Heath et al. [134] focused on computing the affinity factor when the trust relationship is between two individuals, and the expertise and experience factors when the trust relationship is between an individual and a certain topic.

To compute the expertise (or credibility) factor, Revyu tags are inspected. For each tag, all items tagged with that tag are obtained. Then for each item, the mean item rating is obtained and each review of the item is inspected. At the end, each reviewer will have its credibility score updated.

To compute the experience (or usage) factor, Revyu tags and user tags on del.icio.us are inspected. The algorithm counts how many times each reviewer has reviewed a tagged item. At the end, each reviewer will have his tag counts (usage scores) updated.

Note that both algorithms above have one crucial problem: if the user is the only reviewer then this reviewer will obtain credibility and usage scores of 1, which is a full score representing maximum credibility/usage!

Finally, the affinity between two individuals is computed based on the analysis of their reviews in Revyu and some further basic user details from FOAF. The algorithm looks for the items that both reviewers have reviewed. An 'item overlap ratio' is obtained by dividing the number of items reviewed by both peers by the highest number of reviews by either peers. Then, the 'mean rating overlap' is obtained by taking into consideration the average rating distance of both reviewers (the difference in their ratings of each common item they have reviewed). However, how the item overlap ratio is combined with the mean rating distance to obtain the affinity factor has no clear answer yet, although two possible options are provided.

4.2.6 A NodeRank Algorithm for Computing Reputation

The basic goal of Pujol et al. [228] is to use a ranking algorithm to establish the reputation of nodes in a social network. The idea is that properties about a person's degree of expertise (or his reputation) may be inferred from how well this person is connected in his social network.

Pujol et al. [228] build their social networks using information from personal web pages, reports or documents authorship, participation in a project, hierarchical structure in the community or organization, sharing of physical resources, sharing of virtual resources (e.g. news groups, forums, etc.), and email traffic.

A NodeRanking algorithm is proposed for creating a ranking of reputation ratings. The idea is that the ranking of a node would rely on its 'degree of authority', or what may be seen as the degree of 'importance' of the node. "Authority of a node a is calculated as a function of the total measure of authority present in the network and the authority of the nodes pointing to node a."

Note that this method requires no user feedback and is similar to Pagerank. However, while Pagerank uses global information of the network graph, NodeRank uses only local information.

4.2.7 Propagation of Trust in Social Networks

Golbeck and Hendler [114] focus on three properties of trust: transitivity, asymmetry, and personalization. Transitivity implies that trust may be passed between people. For example, if our trusted friend trusts a certain plumber, then we might take our friend's trust into consideration and decide that we will also trust the plumber. However, trust is not strictly transitive. In other words, it is not always the case that if Alice trusts Bob, and Bob trusts Chuck, then Alice should trust Chuck. Asymmetry implies trust is not necessarily reciprocal. In other words, if Alice trusts Bob, then this does not necessarily imply that Bob trusts Alice as well. Finally, personalization implies that trust is 'inherently a personal opinion'. In other words, given the trust values of each node in a social network, calculating the trustworthiness of the node in question will have different results when different peers are performing this calculation.

In this model, nodes label their neighbours as either trusted or not, using the binary values 0,1. When questioning the trust of a given peer, the peer will poll its trusted neighbours only. When the neighbours reply, the peer will average their results and round the final value to obtain a binary value 0,1. Each of the neighbours uses the same algorithm in obtaining their trust value of the peer in question. A variant of this model only rounds the value at the final step when the initial requester receives all results. This is called the non-rounding algorithm, while the first is called the rounding algorithm. Results show that the rounding algorithm outperforms the non-rounding one. This is because rounding intermediate results increases accuracy by removing more error.

4.2.8 Searching Social Networks via Referrals

Yu and Singh [315] focus on using referrals for searching dynamic social networks. However, no social network analysis has been used in the peer selection process. The authors argue that because building and maintaining a peer's social network is not feasible, distributed search through referrals is more promising.

In a referral system, peers send queries to other peers of the selected contacts. A response can either be an answer or another referral, allowing referrals to propagate in the social network. But how do peers select which peers to send their query to? To achieve that, each peer maintains a profile for itself and an acquaintance model for each of its acquaintances, modelled via the vector space model (VSM). Similarly, a query is modelled as a term vector. The similarity between a

query and expertise of a given peer is defined as the cosine of the angle between them. This similarity measure is then used by a peer to help it select other peers to query. In addition to the similarity measure, a peer also considers the sociability of others (or their ability to give good referrals).

Weighted referral graphs are used to help the peer decide which querying peer to follow first. When another referral is received, the peer might decide to pursue it, even if the referred peer is not in its acquaintance list. This is how acquaintances are added. But if an answer is received, the peer updates its acquaintances by assigning rewards and penalties. Both the expertise and sociability vectors of the peer who sent the reply will be updated according to well defined functions.

Note that peers are assumed to be trustworthy is answering queries. For instance it is assumed that a peer answers a query only if it is confident of its expertise, or it may send back another referral only if it is confident in the relevance of the peer being referred.

4.2.9 Reciprocity as a Supplement to Trust and Reputation

An interesting overview to the 'evolution of cooperation' is provided by Mui [207]. In the context of LiquidPub, these theories may be used for defining the various incentives for peers to cooperate or defect (such as lying when rating others or providing unreliable reviews). Four theories are presented by Mui [207]: the group selection theory, the kinship theory, the reciprocation theory, and the social learning theory.

The group selection theory states that cooperation amongst individuals is not consistent with Darwin's theory. As an alternative, it suggests that cooperation is a result of natural selection being based on the group level: species, community, etc. However, strife is often observed within species. This weakens the group theory, suggesting the need for alternative theories.

The kinship theory states that individuals are ready to cooperate and sacrifice themselves for their kins, e.g. parents and children. The closer the relationship with the kin, the more altruism and less aggression is shown, even if it was over one's own personal benefits. However, the main problem with this theory is that it does not answer the relatedness and competition that sometimes exist among kins.

The reciprocation theory states that 'reciprocal altruism' is the reason behind individuals sacrificing their personal gain for the good of others. As such, all cooperative behaviours are viewed as postponing immediate personal gain to benefit from future reciprocations by others. Well defined equations are presented for computing reciprocation. Furthermore, this is the only theory (amongst the presented four) that has been verified by experimental work, proving that 'image scoring does play a role in actual human cooperation'.

Finally, the theory of social learning suggests that in the absence of kinship and reciprocation, cooperation is based on 'cultural transmission'. This implies that individuals learn the most dominant behaviour in their community. However, no experimental work has either proved or disproved this theory. As for the computational trust model, Mui [207] introduces the concept of reciprocity into its proposed computational model. While trust is defined as 'a subjective expectation an agent has about another's future behaviour' and reputation as the 'perception that an agent has of another's intentions and norms', reciprocity is defined as the 'mutual exchange of deeds, such as favour or revenge'. The relation between these three concepts is defined. In short,

increasing reputation results in increasing trust, increasing trust results in increasing reciprocity, and increasing reciprocity results in increasing reputation. Similarly, decreasing any of these three values results in the reverse effect.

4.2.10 Information Based Reputation

[261] illustrates how reputation may be computed by aggregating peer opinions. The proposed method makes use of social network analysis to decide how opinions may be aggregated when there is some sort of dependence between opinions.

The basic idea is that a peer may express its opinion about a given aspect of the object being analysed in a given context. The opinion is described as a probability distribution over an evaluation space E. After forming an opinion, an agent may then share this opinion with the rest of the group. A group discussion then takes place, after which the agent may or may not revise their opinions, based on how much can others convince them about changing their opinion. Finally, the group as a whole tries to reach a unified group opinion. Reputation is then defined as the "social evaluation by the group".

Of course, sharing opinions is the central point of this paper. The idea is that shared opinions affect the future opinions of other peers. Hence, [261] proposes a simple and basic communication model for sharing opinions. Opinions then influence each other based on the "semantic similarity" between the two concepts being evaluated.

Forming opinions is affected by several measures, such as the time decay factor, the reliability of the information source, etc. The accuracy of an agent's opinion may sometime be verifiable within a reasonable amount of time. For example, if one gives their opinion about the weather tomorrow, then this opinion may be verified the next day by actually observing the weather and comparing it to what the agent predicted. However, not all opinions may be verified, such as the opinion about the quality of a given scientific paper. In such cases, the opinion may then be evaluated by comparing it to the group opinion.

Three different methods are then proposed for the aggregation of opinions, in the hope of reaching one final group opinion. The dependent method aggregates the opinions of agents that have been discussing/sharing their opinions together. If this method fails to return results, then the data is deemed inconsistent and the agents should have further discussions or agree to disagree. Otherwise, the $\mathcal{Y}$ method proposed the group opinion with maximum certainty. If this result is rejected by the agents, then the final proposal is to say that the group opinion lies some where between the result computed by the dependent method and that of the independent method, which assumes that the priors are completely independent.

Social network analysis is used to help analyse the dependence between two opinions. For example, in the case of scientific publications, this dependence may be a measure of co-authorship or affiliation. Also, social network analysis may be used to calculate the initial confidence of an agent's opinion. This confidence is the direct result of the agent's expertise in the area it is giving an opinion on. Again, in the field of scientific publications, the expertise may be calculated based on how many highly cited papers is the agent an author of, how many prestigious papers did it review, whether it has a central role in the college, etc.

4.3 Conclusion

Obtaining a fair credit attribution is crucial to the success of the LiquidPub system. Similar to available reputation mechanisms, we plan to make use of both opinions (or reviews) and citations for measuring reputation. Hence, a brief literature review of available similar techniques (such as eBay and Amazon's rating systems and Google's PageRank algorithm) has been briefly presented by the first section of this chapter. However, we also plan to make use of social network analysis to detect redundant ratings, consistent biases, deliberately distorted ratings, unreliable reviewers, etc. Hence, the second part of this chapter has focused on the available mechanisms that makes use of social network analysis for computing notions of trust and reputation.

Part 5

Process Models, Copyright and Licensing Models

5.1 Introduction

The advent of the Web 2.0 or the Social Web has triggered a profound transformation of the process of knowledge production and dissemination: the creation, maintenance, control and sharing of information has become a decentralised, i.e. distributed, collaborative and peer-monitored, process involving global networks of both professionals and volunteers. This has provoked, at least in the last 30 years, a major change in the institutional framework and the social practices that are associated with the production and dissemination of information that has lead today to an "information networked economy" (see Benkler [18]) whose main rules and principles are wholly different from those of traditional economies.

Information is a non-rival resource, that is, consumption by one person does not make it any less available for the consumption of another. Note that this is not a property of new, networked means of information production: it is an exception that exists from the onset in the market of informational and cultural goods: a new reader of Shakespeare's sonnets does not harm any previous reader or prevents more readers to access it. The specificity of information markets may be expressed economically in terms of marginal costs: the increased production of new products does not affect, or poorly affects, the marginal costs of the production. Hence, an exceptional legal apparatus of protection of information and culture markets, in terms of property rights and copyright, has always been necessary in order to assure control and concentration of means of productions for the actors of information markets. This ensemble of legal techniques of control has different histories in the three main domains concerned with the market of ideas: science, authorship and patenting [23]. While these three forms of protection of intellectual creativity were clearly distinguished in history, today some distinctions have been blurred and we are facing a very confused legal framework, whose effects become often harmful for the actors it is supposed to protect [127].

Many studies (for example, Lerner [181]) reveal that the impact of intellectual property of patents on innovation is fairly limited: information is both the input and the output of its own production process (the famous "on the shoulders of giants" effect), that is, new information goods or innovation builds on existing information. The increase of patent protections in the last 150 years

thus increases the cost for the current innovators to access existing knowledge, hence decreasing their potential of creativity. This effect is oddly true today also for the access of scientific knowledge, even if, historically, the status of scientific authorship and that of the inventor were clearly distinguished. Scientific authorship is different from other forms of authorship because it does not result in any property rights for the author: a scientific discovery cannot be copyrighted by an author: historically, the author may claim only the "priority" of his or her discovery and, once the discovery has been peer-reviewed by the community, share it in the "public domain". Now, today, the domain of research papers is not public, but regulated by copyright privileges credited to the publishing companies. One could argue that this could have effects similar to those cited in the world of patenting, that is, that reduced access to previous information results in reduced creativity and originality. However, the comparison with patenting has to be qualified: a lot of knowledge is (and has always been in earlier times) also "hidden" inside companies, who do not disclose it for good, because it is the basis of their business and they use it as a competitive advantage. What has probably changed is that nowadays companies tend to have such knowledge patented much faster (or in an earlier state) that in former times.

Be it as it may, with the advent of the Social Web, the notion of "public domain" as a space for sharing scientific knowledge has re-emerged in the discussion. Still, today the space of knowledge sharing is not a res nullius, as the public domain used to be, but an organized informational space, whose networked capacities have to be defended as "common goods" of the humanity. Hence, we are facing a transition today from a conception of free circulation of ideas in the empty space of the public domain, to a more cooperative idea of sharing knowledge as a common good through a common resource whose use must be regulated, that is, the Web.

This part consists of two sections that assess the impact of the Social Revolution of the World Wide Web both on the production and the distribution of scientific knowledge, with a special focus on intellectual property rights. Section 5.2 provides an overview on the different innovative features and services of the Web 2.0 and gauges its potential for scientific research. First we will review the main social applications that are constitutive of the Social Web, applications that enable interaction, collaboration and sharing between users. We will focus on blogging, podcasting, collaborative content (for example, Wikipedia), social networking (MySpace, Facebook), multimedia sharing (Flickr, YouTube), and social tagging (Deli.cio.us). Our aim is to identify functionalities and applications that can enhance the research process; in particular, we will concentrate on tools for collaborative writing, brainstorming, bibliography and data set sharing, reputation attribution and social networking. Section 5.3 reviews and differentiates the notions of copyright, scientific authorship, and e-commons. The latter is defined as comprising both works that are in the public domain and works that are "copylefted", i.e. licensed by their authors to be copied, shared or modified by anyone provided the copies or modified versions are distributed under the same terms. Different free public licenses will be presented, both copyleft licenses and the copyright licenses provided by Creative Commons. After discussing the relationship between scientific research and the e-commons, we will describe the changes of the copyright and licensing practices in the scientific publishing industry. An overview of the main Open Access models concludes this section.

The whole first part of this section is based on research by Giuseppe Veltri [299]; Sections 5.2 to 5.3.2 are by Luc Schneider, with contributions from Gloria Origgi and Roberto Casati. Section 5.3.3 on copyright and licensing practices, as well as business models, in the scientific publishing industry is by Diego Ponte and Ralf Gerstner.

Figure 5.1: The 'Conversation Prism' in the Web 2.0, by Brian Solis (see `http://www.briansolis.com/2008/08/introducing-conversation-prism.html`).

5.2 The impact of the Web 2.0 on the Research Process

The uptake of social computing applications has been impressive. Social computing applications are defined as applications that enable interaction, collaboration and sharing between users. They include applications for blogging, podcasting, collaborative content (for example, Wikipedia), social networking (MySpace, Facebook), multimedia sharing (Flickr, YouTube), social tagging (Deli.cio.us) and social gaming (Second Life). For a graphical representation of the range of applications involved in social computing see Figure 5.1.

The importance of social computing has been acknowledged the business community, the academic community and by the public opinion at large. It is considered to be a potentially disruptive 'Information Society' development, in which users play an increasingly influential role in the way products and services are shaped and used. This may have important social and economic impacts on all aspects of society. There is, however, little scientific evidence on the take-up and impact of social computing applications. Our objective is to provide a systematic assessment of the principles and potential research uses of social computing applications.

5.2.1 An inventory of the Social Web and Social Computing

This section reviews two important technologies in Social Web: social networks and social computing.

Social networks

Social network sites represent a fundamental layer in the complex phenomenon of the Social Web. We define social network sites (SNSs from now on) as web-based services that allow individuals to

1. Construct a public or semi-public profile within a bounded system,
2. Articulate a list of other users with whom they share a connection,
3. View their list of connections and those made by others within the system.

The nature and nomenclature of these connections may vary from site to site.

While we use the term ”social network site” (such as Friendster, MySpace, LinkedIn and Facebook) to describe this phenomenon, the term ”social networking sites” also appears in public discourse, and the two terms are often used interchangeably. We chose not to employ the term ”networking” for two reasons: emphasis and scope. ”Networking” emphasises relationship initiation, often between strangers. However, what makes social network sites unique is not that they allow individuals to meet strangers, but rather that they enable users to articulate and make visible their ”offline” social networks, or ”latent ties” [133].

While SNSs have implemented a wide variety of technical features, their backbone consists of visible profiles that display an articulated list of Friends who are also users of the system. Profiles are unique pages where one can ”type oneself into being” [277]. After joining an SNS, an individual is asked to fill out forms containing a series of questions. The profile is generated using the answers to these questions, which typically include descriptors such as age, location, interests, and an ”about me” section. Most sites also encourage users to upload a profile photo. Some sites allow users to enhance their profiles by adding multimedia content or modifying their profiles’ look and feel. Others, such as Facebook, allow users to add modules (”Applications”) that enhance their profile.

The visibility of a profile varies by site and according to user discretion. Structural variations around visibility and access are one of the primary ways that SNSs differentiate themselves from each other. After joining a social network site, users are prompted to identify others in the system with whom they have a relationship. Most SNSs require bi-directional confirmation for Friendship, but some do not. The term ”Friends” can be misleading, because the connection does not necessarily mean friendship in the everyday vernacular sense, and the reasons people connect are varied.

The public display of connections is a crucial component of SNSs. The Friends list contains links to each Friend’s profile, enabling viewers to traverse the network graph by clicking through the Friends lists. On most sites, the list of Friends is visible to anyone who is permitted to view the profile, although there are exceptions.

Most SNSs also provide a mechanism for users to leave messages on their Friends' profiles. This feature typically involves leaving "comments," although sites employ various labels for this feature. In addition, SNSs often have a private messaging feature similar to webmail. While both private messages and comments are popular on most of the major SNSs, they are not universally available.

Beyond profiles, Friends, comments, and private messaging, SNSs vary greatly in their features and user base. Some have photo-sharing or video-sharing capabilities; others have built-in blogging and instant messaging technology. There are mobile-specific SNSs (e.g., Dodgeball), but some web-based SNSs also support limited mobile interactions (e.g., Facebook, MySpace, and Cyworld). Many SNSs target people from specific geographical regions or linguistic groups, although this does not always determine the site's constituency. Orkut, for example, was launched in the United States with an English-only interface, but Portuguese-speaking Brazilians quickly became the dominant user group. Some sites are designed with specific ethnic, religious, sexual orientation, political, or other identity-driven categories in mind. There are even SNSs for dogs (Dogster) and cats (Catster), although their owners must manage their profiles.

While SNSs are often designed to be widely accessible, many attract homogeneous populations initially, so it is not uncommon to find groups using sites to segregate themselves by nationality, age, educational level, or other factors that typically segment society even if that was not the intention of the designers.

Currently, there are no reliable data regarding how many people use SNSs, although marketing research indicates that SNSs are growing in popularity worldwide [63]. The rise of SNSs indicates a shift in the organization of online communities. While websites dedicated to communities of interest still exist and prosper, SNSs are primarily organized around people, not interests. Early public online communities such as Usenet and public discussion forums were structured by topics or according to topical hierarchies, but social network sites are structured as personal (or "egocentric") networks, with the individual at the centre of their own community. This more accurately mirrors non-mediated social structures, where "the world is composed of networks, not groups" [310]. The introduction of SNS features has introduced a new organizational framework for online communities, and with it, a vibrant new research context.

Content Creation in Social Computing

The winning principle: harnessing collective intelligence

The central principle behind the success of the giants born in the Web 1.0 era who have survived to lead the Web 2.0 era appears to be this, that they have embraced the power of the web to harness collective intelligence.

As users add new content, and new sites, it is bound in to the structure of the web by other users discovering the content and linking to it. Much as synapses form in the brain, with associations becoming stronger through repetition or intensity, the web of connections grows organically as an output of the collective activity of all web users.

Yahoo!, the first great Internet success story, was born as a catalogue, or directory of links, an aggregation of the best work of thousands, then millions of web users. Google's breakthrough in search, which quickly made it the undisputed search market leader, was PageRank, a method

of using the link structure of the web rather than just the characteristics of documents to provide better search results. eBay's product is the collective activity of all its users; like the web itself, eBay grows organically in response to user activity, and the company's role is as an enabler of a context in which that user activity can happen. Even more, eBay's competitive advantage comes almost entirely from the critical mass of buyers and sellers, which makes any new entrant offering similar services significantly less attractive.

Amazon sells the same products as competitors such as Barnesandnoble.com, and they receive the same product descriptions, cover images, and editorial content from their vendors. But Amazon has made a science of user engagement. They have an order of magnitude more user reviews, invitations to participate in varied ways on virtually every page–and even more importantly, they use user activity to produce better search results. While a Barnesandnoble.com search is likely to lead with the company's own products, or sponsored results, Amazon always leads with "most popular", a real-time computation based not only on sales but other factors that Amazon insiders call the "flow" around products. With an order of magnitude more user participation, it is no surprise that Amazon's sales also outpace competitors. Now, innovative companies that pick up on this insight and perhaps extend it even further, are making their mark on the web.

Wikipedia, an online encyclopaedia based on the unlikely notion that an entry can be added by any web user, and edited by any other, is a radical experiment in trust, applying Eric Raymond's dictum (originally coined in the context of open source software) that "with enough eyeballs, all bugs are shallow," to content creation. Nowadays, Wikipedia is definitely one of the most popular and used web sites. This is a profound change in the dynamics of content creation!

Sites like del.icio.us and Flickr, two companies that have received a great deal of attention of late, have pioneered a concept that some people call "folksonomy" (in contrast to taxonomy), a style of collaborative categorization of sites using freely chosen keywords, often referred to as tags. Tagging allows for the kind of multiple, overlapping associations that the brain itself uses, rather than rigid categories. In the canonical example, a Flickr photo of a puppy might be tagged both "puppy" and "cute"–allowing for retrieval along natural axes generated user activity.

Collaborative spam filtering products like Cloudmark aggregate the individual decisions of email users about what is and is not spam, outperforming systems that rely on analysis of the messages themselves. It is a truism that the greatest Internet success stories do not advertise their products. Their adoption is driven by "viral marketing"–that is, recommendations propagating directly from one user to another. You can almost make the case that if a site or product relies on advertising to get the word out, it is not Web 2.0.

Even much of the infrastructure of the web–including the Linux, Apache, MySQL, and Perl, PHP, or Python code involved in most web servers–relies on the peer-production methods of open source, in themselves an instance of collective, net-enabled intelligence. There are more than 100,000 open source software projects listed on SourceForge.net. Anyone can add a project, anyone can download and use the code, and new projects migrate from the edges to the centre as a result of users putting them to work, an organic software adoption process relying almost entirely on viral marketing.

Main innovative features of Social Computing

Blogging. One of the most highly touted features of the Web 2.0 era is the rise of blogging. Personal home pages have been around since the early days of the web, and the personal diary and daily opinion column around much longer than that, so just what is the fuss all about?

At its most basic, a blog is just a personal home page in diary format. But as Rich Skrenta notes, the chronological organization of a blog "seems like a trivial difference, but it drives an entirely different delivery, advertising and value chain." One of the things that has made a difference is a technology called RSS. RSS is the most significant advance in the fundamental architecture of the web since early hackers realized that CGI could be used to create database-backed websites. RSS allows someone to link not just to a page, but to subscribe to it, with notification every time that page changes. Skrenta calls this "the incremental web." Others call it the "live web".

Now, of course, "dynamic websites" (i.e., database-backed sites with dynamically generated content) replaced static web pages well over ten years ago. What is dynamic about the live web are not just the pages, but the links. A link to a weblog is expected to point to a perennially changing page, with "permalinks" for any individual entry, and notification for each change. An RSS feed is thus a much stronger link than, say a bookmark or a link to a single page.

The number of blogs has doubled every 5-7 months for the last 3 years. Worldwide, in absolute numbers, in October 2006, the specialised blog search engine Technorati (created by Dave Sifry) was tracking over 50 million Blogs. The number increased to 70 million blogs in April 2007 [284]. 120,000 new blogs are created daily - that is about 1.4 blogs created every second of every day. According to Technorati, in October 2008, this figure went up to more than 133 million blogs.

A mapping of the distribution of blogs by language could give an indication of the relative sizes of some individual language-blogospheres. For instance, the Japanese-language blogosphere leads with 37 percent (up from 33 percent in the third quarter of 2006) of the posts, followed closely by the English-language blogosphere at 36 percent (down from 39 percent in the third quarter of 2006). There has been slight decrease in the number of English-language posts (33 percent in March 2007 from 36 percent in October 2006). The Italian-language blogosphere has overtaken the Spanish as the 4th largest. The newcomer to the top 10 languages is Farsi, ranked as the 10th.

Counting blogs based on the country of origin is difficult due to the worldwide phenomenon of people using Anglo-Saxon (US and UK) blogging hosts. A study, Hurst, M., Siegler, M., Glance, N. [143], puts forward a comparison between the geographical location of bloggers and the language in which the blogs are written. While almost 40 percent of blogs are written in English (according to Technorati), some 42 percent of the bloggers claim a location in an English-speaking country. Likewise, 38 percent of the bloggers claim a Chinese location, while only 10 percent of the blogs are written in Chinese.

Podcasting. A podcast can mean either the content itself or the method by which the content is distributed; the latter is also termed podcasting. Podcasts are produced either by 'professional' podcasters or 'private' podcasters (i.e. podcasts created by people, such as bloggers and individual podcasters) and an increasing number of uses are being found for podcasts. In this research, we refer to both podcast content and method. The number of podcasts is difficult to estimate. According to IDATE research released in July 2007, the estimated number of podcasts to date is over 100,000, when only three years ago, there were fewer than 10,000.104 Statistics on the

amount of podcast content and podcast feeds are made available by podcast directories worldwide. Apple iTunes, for instance (see Figure 19), counted over 82,000 podcasts in their directories105 in 2006 (representing a 10 fold increase since 2005).

In terms of the number of podcast feeds, in the US for instance, Feedburner reported more than 40,000 podcast feeds under its management in 2006. In 2006, the creation of podcast feeds averaged 15 percent growth month over month. In August 2007, the figure went up to almost 1 million feeds from more than 500,000 bloggers, podcasters and commercial publishers, currently serving 128,358 podcast feeds (as of 4 August 2007).

In 2008, the Pew Internet and American Life Project found that 19 percent of US Internet users have downloaded a podcast for listening at a future point in time, compared to some 7 in an earlier 2007 survey and 12 percent in following survey in the second half of the same year.

Social Tagging. Tagging describes the act of adding keywords, also known as tags, to any type of digital resource. Tags serve to describe the item and enable a keyword-based classification (knowledge management). They can also be used to search for content. The types of content that can be tagged varies from: blogs (Technorati), books (Amazon), pictures (Flickr), podcasts (Odeo), videos (YouTube), to even tagging of tags. Tags are not only metadata, but also content. Tagging also allows social groups to form around similarities of interests and points of view, hence the term social tagging. Social tagging is one of the Web 2.0 success stories, tapping into the 'wisdom of crowds' - i.e. it lets users connect with others, enabling social discovery and connections. Social tagging leads the way towards a semantic web, in bringing in a meaningful and personal search experience.

There has recently been a dramatic increase in the number of pictures tagged with geographical metadata (a method called geo-tagging or geo-coding). Geo-tagging of photos brings a whole new level of context to images. Flickr's vision on the future of geo-tagging is "show me photos taken within the last 15 minutes within a kilometre of me. In 2006, 2 million photos were geo-tagged in Flickr and users have added, on average, over one million tags per week to the dataset. Flickr allows users to drag photos on to a Yahoo map and mark them with a specific worldwide location. Zooomr is another photo sharing service that provides a geo-tagging tool (Google maps are used instead). As of August 2007, there are 2.6 million geo-tagged photos in Flickr (up from 1.6 million one year ago) [46] In February 2007, Technorati was tracking over 230 million blog posts using tags or categories.

The use of tagging comes in many forms. Photo sharing sites like Flickr allow users to add labels to pictures, and video-sharing sites such as YouTube to tag videos, and Amazon uses tags to classify a product. Google's tagging feature is called "bookmark," though it applies the principles of tagging. Last.fm supports user-end tagging or labelling of artists, albums, and tracks to create a site-wide folksonomy of music. Users can browse via tags, and tag radio to allow users to play music that has been tagged a certain way. The number of bloggers who are using tags is also increasing month on month. About 2.5 million blogs posted at least one tagged post in February 2007. According to Pew Internet and American Life, nearly a third of US Internet users have tagged or categorized content online such as photos, news stories or blog posts in 2006 (Pew Internet, [225]). Some 19 percent of US Internet users watching video online have either rated an online video or posted comments after seeing a video online (Pew Internet and American Life Online Video 2007 [226]).

5.2.2 The Architecture of Participation

In this section we provide a review of the most widely acknowledged principles and dynamics that are shaping the Web 2.0. This constitutes a good starting point to comprehend the nature of the innovations that social computing can bring to knowledge creation and sharing.

"Architecture is politics"

Some systems are designed to encourage participation. In his paper, The Cornucopia of the Commons, Dan Bricklin [39] noted that there are three ways to build a large database. The first, demonstrated by Yahoo!, is to pay people to do it. The second, inspired by lessons from the open source community, is to get volunteers to perform the same task. The Open Directory Project, an open source Yahoo competitor, is the result. But Napster demonstrated a third way. Because Napster set its defaults to automatically serve any music that was downloaded, every user automatically helped to build the value of the shared database. All other P2P file sharing services has followed this same approach.

One of the key lessons of the Web 2.0 era is this: users add value. But only a small percentage of users will go to the trouble of adding value to your application via explicit means. Therefore, Web 2.0 companies set inclusive defaults for aggregating user data and building value as a side effect of ordinary use of the application. As noted above, they build systems that get better the more people use them. Mitch Kapor once noted that "architecture is politics." Participation is intrinsic to Napster, part of its fundamental architecture.

This architectural insight may also be more central to the success of open source software than the more frequently cited appeal to volunteerism. The architecture of the Internet, and the World Wide Web, as well as of open source software projects like Linux, Apache, and Perl, is such that users pursuing their own "selfish" interests build collective value as an automatic by-product. Each of these projects has a small core, well-defined extension mechanisms, and an approach that lets any well-behaved component be added by anyone, growing the outer layers of what Larry Wall, the creator of Perl, refers to as "the onion." In other words, these technologies demonstrate network effects, simply through the way that they have been designed.

These projects can be seen to have a natural architecture of participation. But as Amazon demonstrates, by consistent effort (as well as economic incentives such as the Associates program), it is possible to overlay such architecture on a system that would not normally seem to possess it.

Technical devices enabling user added value

RSS also means that the web browser is not the only means of viewing a web page. RSS is now being used to push not just notices of new blog entries, but also all kinds of data updates, including stock quotes, weather data, and photo availability. But RSS is only part of what makes a weblog different from an ordinary web page. Tom Coates remarks on the significance of the permalink, the device that turned weblogs from an ease-of-publishing phenomenon into a conversational mess of overlapping communities. For the first time it became relatively easy to gesture directly at a highly specific post on someone else's site and talk about it.

In many ways, the combination of RSS and permalinks adds many of the features of NNTP,

the Network News Protocol of the Usenet, onto HTTP, the web protocol. The ”blogosphere” can be thought of as a new, peer-to-peer equivalent to Usenet and bulletin boards, the conversational watering holes of the early Internet. Not only can people subscribe to each others' sites, and easily link to individual comments on a page, but also, via a mechanism known as trackbacks, they can see when anyone else links to their pages, and can respond, either with reciprocal links, or by adding comments.

Interestingly, two-way links were the goal of early hypertext systems like Xanadu. Hypertext purists have celebrated trackbacks as a step towards two way links. But note that trackbacks are not properly two-way–rather, they are really (potentially) symmetrical one-way links that create the effect of two way links. The difference may seem subtle, but in practice it is enormous. Social networking systems like Friendster, Orkut, and LinkedIn, which require acknowledgement by the recipient in order to establish a connection, lack the same scalability as the web. As noted by Caterina Fake, co-founder of the Flickr photo sharing service, attention is only coincidentally reciprocal. (Flickr thus allows users to set watch lists–any user can subscribe to any other user's photostream via RSS. The object of attention is notified, but does not have to approve the connection.)

Blogging as a filter harnessing collective intelligence

If an essential part of Web 2.0 is harnessing collective intelligence, turning the web into a kind of global brain, the blogosphere is the equivalent of constant mental chatter in the forebrain, of conscious thought. And as a reflection of conscious thought and attention, the blogosphere has begun to have a powerful effect. First, because search engines use link structure to help predict useful pages, bloggers, as the most prolific and timely linkers, have a disproportionate role in shaping search engine results. Second, because the blogging community is so highly self-referential, bloggers paying attention to other bloggers magnifies their visibility and power. The ”echo chamber” that critics decry is also an amplifier. If it were merely an amplifier, blogging would be uninteresting. But like Wikipedia, blogging harnesses collective intelligence as a kind of filter. What James Suriowecki calls ”the wisdom of crowds” comes into play, and much as PageRank produces better results than analysis of any individual document, the collective attention of the blogosphere selects for value.

While mainstream media may see individual blogs as competitors, what is really unnerving is that the competition is with the blogosphere as a whole. This is not just a competition between sites, but a competition between business models. The world of Web 2.0 is also the world of what Dan Gillmor calls ”we, the media,” a world in which ”the former audience”, not a few people in a back room, decides what is important.

Patterns of participation

In order to understand social computing adoption, there is a need to see how people approach these technologies. Social computing is used not only by the few people posting blog entries, photos on Flickr and videos on YouTube, but by a large share of Internet users in many different ways. The present research confirms that, statistically, the pattern of participation in social computing follows what has been described as a power law distribution (R. Mayfield based on `http://`

www.orgnet.com/BuildingNetworks.pdf).

Moreover, the behaviour of "passive users" is increasingly being explored via technological means. Simply reading or using social computing content can leave traces which can be used (anonymously) as a way of sharing preferences and interests (practically 100 percent of Internet users). The intensity of online participation then diminishes gradually (as described by the Concentric Model of Participation Intensity (CPMI) to at least a third (30-40 percent) of Internet users using social computing content e. g. reading blogs, or watching user-generated videos on YouTube, listening to podcasts, visiting wiki sites, or visiting/using social networking sites. Some 10 percent of Internet users provide feedback (posting comments on blogs and reviews) or share content on Flickr, or YouTube, or tag content in deli.cio.us. Only around 3 percent of Internet users in Europe are "creators" e.g. they create blogs or Wikipedia articles, or upload their user-generated videos on YouTube or photos on Flickr.

People also switch between activities. For example, while reading blogs, they may also visit social networking sites, contribute to Wikipedia, or upload their photos on Flickr. The latest surveys from Forrester (see Figure 10) show that the so-called 'joiners' (representing, according to Forrester, about 20 percent of US adult online population and mostly comprising Generation 'Y' i.e. 18-25 year olds) do a variety of online activities. For example, apart from using social networking sites, 56 percent of them also read blogs, while 30 percent publish blogs.

Another important aspect of social computing is the move from an 'in group' dimension of use and computing to an 'out group' one. The developing of web applications that are designed to expand the range of collaboration is one of the main features of social computing and Web 2.0. The move is from an 'in-group', based on the peers locally available, to a 'out-group' dimension that allows cooperation with individuals that are not immediately part of our environment.

Conclusion

Collaboration is not strictly defined in a top-down process, setting a team and inviting individuals from already known work environments. Instead, a bottom-up process is central in many applications of the Web 2.0, in which people are 'pulled' towards projects or groups by common interests and aims. In this case, the structure of groups is fluid an in constant change, size can be large and collaboration is structured so that the raw power of big numbers can exploited, usually dividing large and complex tasks in small ones.

5.2.3 The Social Web and Research Process

Applying the potential of the Web 2.0 to the research process

A number of functionalities were identified as crucial in applying the potential of the Web 2.0 to the research process by a panel of researchers at the Institute Nicod (CNRS) between September and November 2008. These were:

1. to edit a document individually or in a group, in real time or not;
2. to share with selected users or with the wider community of web users (the in group/out

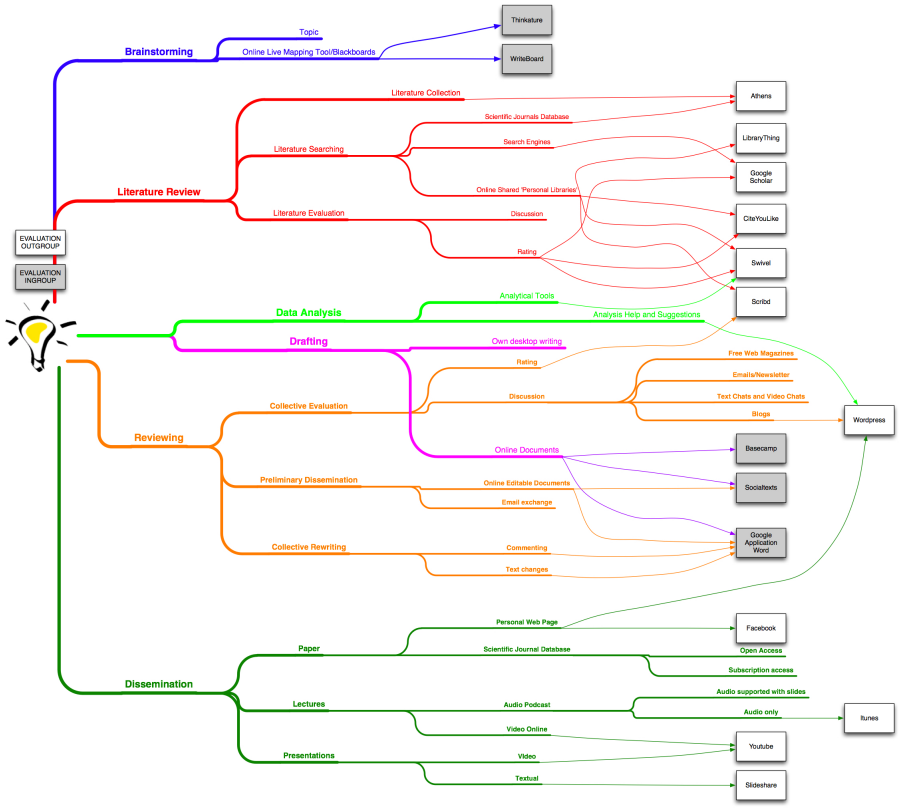

Figure 5.2: Diagram of available tools within the different steps of the research process.

group dimension) integrating with existing social networks websites (CiteULike, Delicious, etc.);

3. to evaluate a document through commenting, ranking;

4. to allow categorization through tags and therefore allow retrieval of similar documents;

5. the capacity of uploading files of different format (word, PDF, rtf, etc.);

6. implementation of reputation and history of reviewers;

7. an open and modular systems of add-ons to incentive user-generated application and future functionalities.

From an exploration of currently available tools, as shown in Figure 5.2, there is not an all-in-one tool that has all the 'desiderata' functionalities. Hence, the different steps of the research process can benefit only from different tools at different stages. The sets of functionalities above described are implemented in several Web 2.0 applications.

Overview of functionalities

A first relevant separation is between brainstorming tools and collaborative writing ones. At the moment, there is no Web 2.0 tool that comprehends brainstorming using 'virtual boards', diagrams and other mind-mapping tools with online writing collaborative tools. Web applications such as Thinkature[1], Mindomo[2] or MindMeister[3] do not provide online collaborative writing tools for documents but not only for diagrams, flow charts and mind mapping, but the MarraTech conference system[4] allows for sharing presentations and blackboards. Although these tools are useful at different stages of the drafting and writing process, a complete application will allow users to go back to their 'brainstorming' steps to re-think or re-elaborate about the document under writing process.

This leads us to the wider issue of a better integration between different writing tools such as text documents, spreadsheets, notebooks, slides, etc. Among the tools that allow a good level of integration, Zoho[5] and Google Docs[6] offer rather interesting examples of the shape of things to come, but currently there are no productivity applications suites available that offer such integration (another example is ThinkFree[7], but it is not free to use).

Another relevant separation of functionalities is between collaborative writing tools and Web 2.0 references sharing applications. For example, online applications as Google Docs do not provide any tool for a researcher to build up a relevant personal library and to share it with other members of a team or with the web. There are already available such services Web 2.0 applications, for example CiteULike[8] that is explicitly for academic researchers or LibraryThing[9] aimed to a more general public. CiteULike and Bibsonomy[10] represent a interesting implementation of reference sharing and it allows a social evaluation of academic articles that is very handful when in the process of selecting and looking for papers on a given topic. Both CiteULike and Library Thing not only allow evaluation of papers or books, but they allow users to make recommendations and to access the library of other users that might have similar research or cultural interests.

Section 5.2.2 points us to another important set of functionalities that is represented by the evaluation of a document, a draft or a book by a community of users or readers. Tools such as Scribd[11] or Docstoc[12] are the most common examples of websites constituted by a database of documents uploaded and evaluated by users. A rating and award system is implemented in both tools to produce a user-generated selection and promotion of the most valid documents, in addition texts can be commented and reviewed and shared on other website through links and embedded 'text reader' provided by both websites. Recently, the online suite of applications Zoho introduced

[1] http://thinkature.com
[2] http://www.mindomo.com
[3] http://www.mindmeister.com
[4] http://www.marratech.com/
[5] http://www.zoho.com
[6] http://docs.google.com
[7] http://www.thinkfree.com
[8] http://www.citeulike.org
[9] http://www.librarything.com
[10] http://www.bibsonomy.org/
[11] http://www.scribd.com
[12] http://www.docstoc.com

Zoho Share[13] that essentially replicates Scribd functionalities in the Zoho environment but that it is integrated with Zoho Writer, Zoho Spreadsheet and other applications.

Of particular interest is the possibility to attribute reputation 'credits' to commentators and reviewers as incentives to a wider and regular process of social evaluation and to track commentators and reviewers history. There are few Web 2.0 tool currently available for such task and the most interesting are: Intense Debate[14]; coComment[15] and SezWho[16]. These tools aim to provide a sort of unified profile and history for commentators and reviewers and they want to create a cross community reputation and rating service.

Currently, the main problem is the lack of connection between communities, therefore reviewers and commentators have multiple identities and their reputation is fragmented in different areas. Tools such as IntenseDebate or SezWho are meta-reputational databases that should allow users to retain their history of commenting and reviewing on any social web site they want to interact with. Portability of reputation history and ID are likely to be crucial issues in the developing of the Social Web. A solution to this problem could be the OpenID initiative[17]. See Figure 5.3.

The last functionality that would be desirable to have implemented and that does not have almost any implementation of tools available is dataset sharing. One of the few tools is represented by Swivel[18], a Web 2.0 website that allows datasets sharing and basic statistic manipulations that can be saved and shared with other users. The underlying idea is to let people explore data and 'play' with it so that secondary data analysis can be shared and pursued almost as an 'hobby'. The great limitation of this tool is that datasets from public institutions (such as UN, OECD, etc) are available but there are almost none universities involved. It is designed to involve all community members and does not allow 'private sharing'.

On the contrary, a current undergoing project, 'Dataverse'[19], at Harvard University is targeting the research community. The description of the project is: 'The Project is an open-source software development community, housed at the IQSS. Via web application software, data citation standards, and statistical methods, the Dataverse Network project increases scholarly recognition and distributed control for authors, journals, archives, teachers, and others who produce or organize data; facilitates data access and analysis for researchers and students; and ensures long-term preservation whether or not the data are in the public domain'. This project developed an open source client software that allows the creation of individual 'dataverses' are self-contained virtual data archives, which are served by a Dataverse Network, and appear on the web sites of authors, teachers, journals, granting agencies, research centres, departments, and others. According to the developers of this project 'each dataverse presents a hierarchical organization of data sets, which might include only studies produced by the dataverse creator (such as for an author or research project), those associated with published work (such as replication data sets for journal articles), or data sets collected for a particular community (such as for a journal's replication archive, or a college class or subfield)'.

[13] http://share.zoho.com/
[14] http://www.intensedebate.com
[15] http://www.cocomment.com
[16] http://sezwho.com, recently acquired by JS-Kit http://blog.js-kit.com/2009/03/06/js-kit-acquires-sezwho/
[17] http://openid.net
[18] http://www.swivel.com/
[19] http://thedata.org/

Figure 5.3: Reproduction from `http://sezwho.com/`

Winning options

In conclusion, there are no tools that comprehend all the 'desiderata' functionalities, but we can select few interesting examples of available Web 2.0 applications among those we have described so far that might be a useful starting point:

1. Zoho represents the most complete online collaborative writing suite of tools that includes several different functionalities from an editor as in Google Docs to a public repository such as Scribd.
2. Thinkature is most complete online mind-mapping tools
3. CiteULike is a likeable example of an online shared personal library of references
4. Scribd is an impressive tool for documents sharing, social evaluation and dissemination
5. IntenseDebate is an inspiring example of a basic reputation system across different communities of the Social Web.
6. Facebook can be represent an example of social web tool in which applications are user-generated as add-ons to the basic service introducing new practices of the tool itself.
7. The DataVerse Project and Swivel are a very interesting exploration of datasets sharing and manipulation.

Open issues: Social Web and Sciences

At this point we should discuss a common critique to the approach of social computing and 'wisdom of crowds' applied to the domain of science. The wisdom of crowds depends on the existence of crowds, however, there are three barriers to Social Web extracting the wisdom in sciences as

it does elsewhere. The first is the lack of a crowd - or the 'small world problem'- not only is the total number of scientists in any one field rather low in terms of Internet numbers, but it is even lower in reality with specialization. Some research domains have small numbers for Social Web. For example, there is much more college students for Facebook than there are neuroscientists for a potential "Neurobook".

The second problem is that scientific communication is quite different than normal human communication. Scientists talk to their friends, but when talking to people they do not know, it is much more formal. They use communication to specify theories and to claim ground as theirs. Hence the problem is that people with common knowledge do not share it with each other, simply because of social competition (and time constraints). These issues are known as barriers for knowledge sharing, unwillingness to share one's best ideas and use ideas of others (not-invented-here syndrome) being examples [16, 241]. It is the barrier based on the idea that what does not get shared anyway is not likely to get shared simply because the technology exists to share it. In other words, if a scientist is not going to share something to his/her colleagues at a conference, he/she probably is not going to share it by web means.

The third problem is that there are no rewards for participating in these new forms of communication. The risks associated with sharing and opening up a scientist's work to other peers before his 'paternity' has been acknowledged are present but the rewards are not clear yet. The basic idea is that one reward is constituted but the contributions that will rise by the sharing process but credit attribution, reputation and intellectual property are still unsolved issue that create a formidable obstacle for adopting Web 2.0 tools in sciences.

In conclusion, if we reconsider together these three problems - crowd is too small, communication is too formal, and no one gets rewarded - how do we overcome this to get Social Web's very-real benefits into the sciences? We propose here few starting points more than exhaustive answers to the aforementioned three problems:

1. Increase the size of the crowd. This potential solution starts with Open Access. More people reading the source materials is simply the only possible way to go. We need to abandon the 'Walled Garden' approach to the content. There are people out there who can learn this, but not without access to the canon. It also requires Research Web - that is to say the reformatting of the scholarly canon so that it is not just legally accessible as a set of PDF files, but something that can be endlessly manipulated, searched, indexed, and more. Scientific knowledge is inherently compatible with the idea of wiki - each paper is a nodal set of relationships between linkable entities - but it needs to be reformatted first. At the moment the combination of publisher firewalls and underlying data formats is a "bottleneck point" on Social Web utility, because it keeps anyone who is not already in the Science community on the outskirts.

2. Incentivise participation. This is both a combination of Social and Research Web. It could be as simple as having rewards for whoever creates the most bookmarks, curates the local edges of a semantic graph, tags the most papers. It could be as simple as having a Technorati rank be considered in faculty hiring (though this is as fraught with problems as citations, if not more). It could be asking for proof of the reverberation of one's research and ideas in any number of ways. The point is to get an environment where scientists see value in talking to each other more than they do.

These are only two starting points that have the function of showing how solutions might be at hand to the three main barriers of adopting Web 2.0 tools in the sciences. It is a new challenge but the potential might bring great benefit to science in general and to social sciences in particular. To illustrate an example of such potential benefit is the aim of the next section of this paper, in which we present an example of how the adoption of social computing might change and improve a research practice in the social sciences.

5.3 The Impact of the Web 2.0 on Copyright and Licensing

In the following, we will gauge how the advent of the Social Web, in particular the ease of copying and sharing content on the Web 2.0, have transformed our conception of intellectual property as well as our practices related to the latter. First, we will provide a conceptual analysis and historical overview on the intertwined notions of copyright and what one may call, after Lessig [182, 183] and Boyle [36], the ”e-commons”, which is distinct from the public domain and covers both copylefted and ”free”, i.e. unpropertised, resources. Since patenting is of prime importance for scientific research, this aspect of intellectual property will be also briefly discussed, together with the concept of scientific authorship. Second, we will assay two recent reflections as to the need to limit or at least redefine the scope of intellectual property in the digital era for the sake of protecting the freedom of scientific research. James Boyle [36] criticises what he calls a ”second enclosure movement”, a general tendency in current national and international legislations to fence off and slowly carve up the public domain, which may stifle intellectual and scientific creativity by reducing the ”commons” of freely available results and data. On a different note, Stevan Harnad [126] pleads for a distinction of two dimensions of copyright, namely protection from theft of ideas (plagiarism) and protection from theft of text (piracy) and argues that only the former is relevant for scientific authorship that aims for impact and not for income. Both critical appraisals of the notion of intellectual property aim at the defence of a ”scientific commons” in which authors may self-archive their papers, results and data for every other scientist to use and build upon. Third, we will describe how the copyright and licensing models and the related business models in scientific publishing have been transformed due to the Web 2.0. In particular, we will address the issue of how the scientific publishing industry has started to adapt to the challenges of the Internet and the Social Web by diversifying its licensing and business strategies.

5.3.1 Copyright and the E-Commons

Defining copyright

Copyrights are a kind of intellectual property, the other two categories being patents and trademarks [166]. The rationale of patent law is to protect the exclusive rights as to the exploitation or distribution of inventions, i.e. new products, devices and processes, or improvements thereof, with the explicit exclusion of ideas and methods of operation, e.g. the buttons on a radio [166]. Trademark protection aims at the exclusive right to use a certain product names [166]. The scope of copyright is original expressions [166].

More precisely, the purpose of copyright is to grant the author of an original work exclusive rights for a limited time period with respect to the publication, distribution and adaptation of that

work. After that period time the work enters the public domain [21]. However, most legislations allow for "fair" exceptions to the author's exclusive rights, and giving users certain rights, such as to make copies for private use or to quote from published works, under the condition to give credit to their authors.

Copyright is intended as giving authors control over and profit from their works, thereby encouraging and fostering the creation of new works and the flow of ideas and learning. This seems to be mandatory in an epoch when an increasing number of people earn their living from intellectual achievements [283]. Intellectual property is necessary for an author to be able to make money of his work (ibid.). Contrary to a modern misconception, copyright is essentially a right of authors and creative minds, not of publishing companies (ibid.). The problem of piracy and copyright infringements is not merely loss of income, but also distorted reproduction (ibid.).

Copyright applies to the expression of any idea or piece of information that is sufficiently original. In other words, copyright does not concern ideas or bits of information, but primarily the manner in which they are expressed [166]. As such, a wide range of creative, intellectual, or artistic forms are covered, including news paper articles, poems, scientific papers, academic theses, plays, novels, personal letters, but also movies, dances, musical compositions, recordings, paintings, drawings, sculptures, photographs, software, radio and television and broadcasts.

Evolution of copyright

The history of copyright starts in the 18th century, with a very rich previous history in the XVII century [55]. In fact, copyright law has its origin in the monopolies that appeared with the development of presses: publishers and bookbinders were organized in guilds and protected their primacy in information dissemination by keeping their manufacture methods secret. Indeed, until the early 18th century, publishers hold more rights over printed works than their authors [166]. The Statute of Anne (1710) in Britain can be regarded as the first copyright act; it established both the author of a work and its publisher as owners of the right to copy that work for a period of time of 21 years [166, 21]. The Copyright Clause of the United States Constitution (1787) provided for a legislation that was much more in favour of the authors: "To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries."

In 1886, under the instigation of Association Litteraire et Artistique Internationale (AIAI) and its president, the French poet and novelist Victor Hugo, the Berne Convention first established a form of international recognition of copyrights. It was influenced by the French legal concept of "droit d'auteur" and attributed the exclusive ownership of a work to its author. In the 160 countries currently adhering to the Berne Convention, copyrights for creative works generally are automatically in force as soon as they are written or recorded on some physical medium, unless the author explicitly disclaims them, or until the copyright expires and the work falls into the public domain [98].

The regulations of the Berne Convention have been incorporated into the World Trade Organization's TRIPS agreement (1995), thus giving the Berne Convention effectively near-global application. The 1996 WIPO Copyright Treaty 1996 extended copyright to computer programs [21], while 2002 WIPO Copyright Treaty enacted greater restrictions on the use of technology to copy works in the nations that ratified it.

At present, all member states of the EU are signatories of the Berne convention. Furthermore, in the last decade of the 20th century, numerous steps have been taken to harmonize national legislations regarding copyright. The EC directive on the legal protection of computer programs (91/250/EEC) in 1991 was the first major attempt to harmonize national copyright laws within the European Economic Community. in 1993, a common term of copyright protection, 70 years from the death of the author, was determined by Council Directive 93/98/EEC harmonizing the term of protection of copyright and certain related rights. Since then, harmonization of European copyright law was increased by a number of directives, notably Directive 96/9/EC of the European Parliament and the Council of 11 March 1996 on the Legal Protection of Databases, Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonization of certain aspects of copyright and related rights in the information society, the Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights and the Directive 2006/116/EC of the European Parliament and of the Council of 12 December 2006 on the term of protection of copyright and certain related rights. The latter confirms the term of protection of copyright to 70 already fixed in the Council Directive 93/98/EEC.

Establishing copyright claims

In all countries where the Berne Convention applies, copyright is automatic, and need not be obtained through official registration with any government office. Once an idea has been reduced to tangible form, for example by securing it in a fixed medium (such as a drawing, sheet music, photograph, a videotape, or a computer file), the copyright holder is entitled to enforce his or her exclusive rights. However, in jurisdictions where the laws provide for registration, it serves as legal evidence of a copyright claim. For example, in the USA it is mandatory to register copyrights with the United States Copyright Office before a infringement suit may be filed in court.

In some countries, e.g. in the UK, commercial services provide a registration facility where copies of work can be deposited to establish legal evidence of a copyright claim. For the same purpose, in most countries inside and outside of the European Union, there are also legal requirements to file certain published works with the respective national library, especially if an ISSN or ISBN has been requested for the latter.

Public domain vs. e-commons

Following Boyle [36], we distinguish two distinct domains ”outside” of the area of intellectual property, namely the public domain and the e-commons. Both notions are evasive and also delusively close to each other, so that it is mandatory to spend some space to discuss and compare them. The concept of e-commons is the most important one, as it is tied to the practice of copyleft licensing.

Public Domain. The notion of public domain stems from the French ”domaine public” which made its way into international and national law through the Berne Convention [188, 36]. David Lange [174] was the first to raise the issue of the necessity to delimit and defend the public domain. Lange [174] argues that the very imprecision of the notion of intellectual property is one of the

major reasons for its "reckless expansion"; the remedy is to acknowledge a "'no-man's land' at the boundaries" of intellectual property [174]. However, Lange does not provide a further clarification of the concept of public domain, nor what individual rights exist within it [36].

Lange's article triggered a whole literature on the topic of public domain. Lindberg and Patterson [185], for instance, proposed to view copyright as a set of temporary and constrained privileges that feeds the public domain with works as their copyrights expire. Jessica Litman [188] contends that the main role of the public domain is allowing copyright law to function despite the unrealistic conception of individual creativity it presupposes. She defines the public domain as a "commons that includes those aspects of copyrighted works which copyright does not protect" [188]. That is, according to Litman's definition, the public domain comprises the re-usable unprotected elements in copyrighted works as well as works that are completely unprotected [36].

Yochai Benkler's [17] approach to the evasive notion of public domain is comparatively pragmatical: the public domain is the totality of all uses, works and aspects of works that can be identified as free by lay people without carrying out a sophisticated legal inquiry into individual facts [17]. According to Boyle [36], Benkler's definition is intended to raise the issue whether lay people really have reliable intuitions as to whether a certain resource is free, i.e. both uncontrolled by someone else and free of charge. Boyle (ibid.) takes a contextualist, if not sceptical stance, on this issue: the delimitation of the public domain "depends on why we care about the public domain, on what vision of freedom or creativity we think the public domain stands for, and what danger it protects against" (ibid.). A certain pluralism about the notion of public domain is the consequence (ibid.).

E-Commons. The term "commons" has come to denote "wellsprings of creation that are outside of, or different from, the world of intellectual property", as is for instance regarded the Internet [36]. As such "commons" or "e-commons" and "public domain" would appear to be synonymous. But Larry Lessig [182, 183] proposes a more restrictive definition: "e-commons" is the totality of works or information the uses of which are maybe not necessarily free of charge, but are such as to be unconstrained by the permission or authorisation of somebody else, certain liability rules excepted. A similar delineation of the concept of commons is proposed by Benkler [19]. The focus is on control and the freedom from the will of another [36] rather than on absence of costs: intellectual property should not restrain innovation in form of a monopoly [36].

Hence being in the e-commons is compatible of being owned individually or collectively. A good example is open-source software that is available under so-called "copy-left" licenses that are actually copyright licenses granting end-users the right to modify or copy the software or any other expression of content as long as these uses comply with the copyleft license [36]. We will discuss the notion of copyleft below.

Hence, the distinction between public domain and e-commons is that the first is based on the dichotomy between the domain of property and the domain of the free, while the second draws the dividing line between the domain of individual control and the domain of "distributed creation, management and enterprise" [36]. Not only is the e-commons compatible with constraints, but the successful examples of e-commons, like open source software, actually presuppose constraints, be they legal - in the form of liability rules - or based on shared values and norms and prestige networks (ibid.).

It is important to note that the e-common is ”outside” of the domain of intellectual property not in the sense that it excludes property rights, but only in the sense that it precludes that they may become an obstacle to innovation and intellectual creativity. Copyleft licenses which are the backbone of the e-commons actually exploit intellectually property rights in order to prevent the abuse of the very same rights, as we shall see below. Thus, the e-commons stands squarely on the ground of intellectual property. However, in a more liberal reading, which is adopted by Boyle, as we shall see below, the notion of e-commons covers both resources subject to intellectual property (but are copylefted) and resources that are free in the sense of being part of the public domain stricto sensu. That is, the e-commons in the wider sense includes the public domain, while stretching over into the area of intellectual property.

The Cornucopia of the E-Commons

Defining Copyleft. Copyleft [64] is used to stipulate the copying terms for a work with a license. According to the Free Software Definition[20], the purpose is to give each person detaining a copy of the work the same liberties as the author, in particular:

1. the freedom to use and study the work,
2. the freedom to copy and share the work with others,
3. the freedom to modify the work,
4. the freedom to distribute modified and therefore derivative works.

In order to ensure that a derivative work is distributed under the same terms, and thus to be copyleft, the license has to stipulate that the author of a derived work can only distribute it under the same or an equivalent license. If the license only requests that the author of a derived work distributes it under the same or a *compatible* license, it is a *share-alike* license. Such licenses may impose additional restrictions, such as prohibiting commercial use (as some Creative Commons licenses do, see below).

Thus, copyleft uses copyright law to remove restrictions related to the distribution of copies and modified versions of the protected work, while requiring that these copies and modified versions preserve the same freedom as the original. As opposed to traditional copyright that locks a work up, copyleft prevents the locking up of the work and its derivative works.

Copyleft is applied to computer software, documents, music, and art. Via a copyleft licensing scheme, an author may permit to everyone who receives a copy of his to reproduce, adapt or distribute it under the provision that the copies or adaptations are also licensed under the same copyleft scheme [270]. Thus copyleft can be regarded as an alternative to letting fall a work wholly into the public domain, namely as a copyright licensing scheme under which an author gives up some of his/her exclusive rights as to the reproduction, distribution or adaptation of his/her work (ibid.).

[20] `http://www.gnu.org/philosophy/free-sw.html`

Short history of copyleft. The idea of copyleft licences originated in 1975, when Dennis Allison wrote a specification for a simple version of the BASIC programming language, Tiny BASIC [4]. This specification appeared in Dr. Dobb's Journal of Tiny BASIC, which computer hobbyists used to write their own BASIC interpreters to be published in the same journal [306].

In 1984, Richard Stallman decided to create what was to be the first real copyleft license, the Emacs General Public License [82], the first copyleft license, after having become irritated by the fact that the company Symbolics, which he supplied with a public domain version of a Lisp interpreter he was working at, refused to allow him access to the changes made by company to the original product. The Emacs General Public License was to develop into the GNU General Public License [271].

The GNU General Public License[21] was designed to make sure that the source-code remained open and freely available in order to foster the sharing of ideas, but it did not exclude commercial usage [21]. The most successful GNU project was Linux, that was started in the 1991 by Linus Torvalds [271], followed by Wikipedia (wikipedia.org), an online collective encyclopaedia that is collaboratively maintained by millions of volunteers, and which is licensed under the GNU GNU Free Documentation License (GFDL) [21].

These projects have inspired Lawrence Lessig and others to establish Creative Commons[22] in 2001 with the support of the Centre for the Public Domain. Creative Commons is an organisation which offers a wealth of copyright licenses, ranging from public domain licenses to sampling licenses, all with the aim of encouraging creative freedom. Releasing work under a Creative Commons license is not the same as giving it away, but it licenses 'reuse' under the conditions defined by the licence chosen by the author [21]. Creative Commons licenses are offered, together with the "all rights reserved" model of traditional copyright, in Knol (knol.google.com), an online knowledge resource provided by Google as an alternative to Wikipedia.

Open Source copyleft licenses. The classical example of copyleft licenses are the GNU General Public License[23] (GPL) and the GNU Lesser General Public License (LGPL) [24]. The GPL stipulates that derivative or linked works must also use the GPL or a compatible license. The LGPL requests direct derivatives of the work to be released under LGPL or a compatible license, but allows any code under any license to link to the LGPL-licensed code. A typical example is that of a library which would be incorporated into a larger work. The LGPL is now generally deprecated by its originator, the Free Software Foundation.

A lesser used alternative for GNU is the GNU Affero General Public License (AGPL) [25]. Similar to the GPL, this license requires source-code modifications to be published if the software itself is distributed, and also if it is used to provide a service via a network (for example, as an Internet application, which runs on the host company's server and thus is not "distributed" in the terminology of the GPL). Finally, the GNU Free Documentation License [26] is a copyleft license designed for textbooks and manuals.

[21] http://www.gnu.org/licenses/gpl.html
[22] http://creativecommons.org
[23] http://www.gnu.org/licenses/gpl.html
[24] http://www.gnu.org/licenses/lgpl.html
[25] http://www.gnu.org/licenses/agpl.html
[26] http://www.gnu.org/licenses/fdl.html

Note that not all open source licenses are copyleft licenses. Some are permissive licenses that offer many of the same freedoms as releasing a work and letting it fall into public domain. For example, the Berkeley Software Distribution (BSD) license [27] allows anyone to do whatever they wish with the code as long as they reproduce the original copyright notice.

Creative Commons licenses. An alternative to the GNU licenses is the suite of copyright licenses provided by Creative Commons[28]. Creative Commons copyright licenses have been ported to over 45 international jurisdictions and by the year 2008, about 130 million works have been licensed under a Creative Commons scheme. The current version of the Creative Common licenses is 3.0 [29].

The generic Creative Commons licenses are designed to be jurisdiction-neutral, but to some extent are founded upon the U.S. Copyright Act. This makes it sometimes necessary to align theses licenses with other national legislations. Therefore, the Creative Commons model has three layers: the human-readable Commons Deed, the lawyer-readable Legal Code, and the machine-readable Digital Code or metadata. With the support of an international network of legal experts, Creative Commons seeks to port the Legal Code to a particular jurisdiction, while the Commons Deed and Digital Code always remain the same (see Figure 5.4).

Creative Commons provides the following license conditions as options to the licensor:

1. Attribution: the licensor allows others to copy, distribute, display and perform his/her copyrighted work as well as any derivative work based on it provided credit is given in the manner requested by the licensor.
2. Share-alike: the licensor allows others to distribute derivative works only under the same license that governs his/her copyrighted work.
3. Non-commercial: the licensor allows others to copy, distribute, display and perform his/her copyrighted work as well as any derivative work based on it only for non-commercial purposes.
4. No Derivative Works: the licensor allows others to copy, distribute, display and perform exclusively verbatim copies of his/her copyrighted work, but no derivative works based on the latter.

The Creative Commons licenses combine the aforementioned license conditions:

1. Attribution (by) [30]
2. Attribution ShareAlike (by-sa) [31]
3. Attribution No-Derivatives (by-nd) [32]

[27] http://www.opensource.org/licenses/bsd-license.php
[28] http://creativecommons.org/about/licenses
[29] http://creativecommons.org/about/history
[30] http://creativecommons.org/licenses/by/3.0/legalcode
[31] http://creativecommons.org/licenses/by-sa/3.0/legalcode
[32] http://creativecommons.org/licenses/by-nd/3.0/legalcode

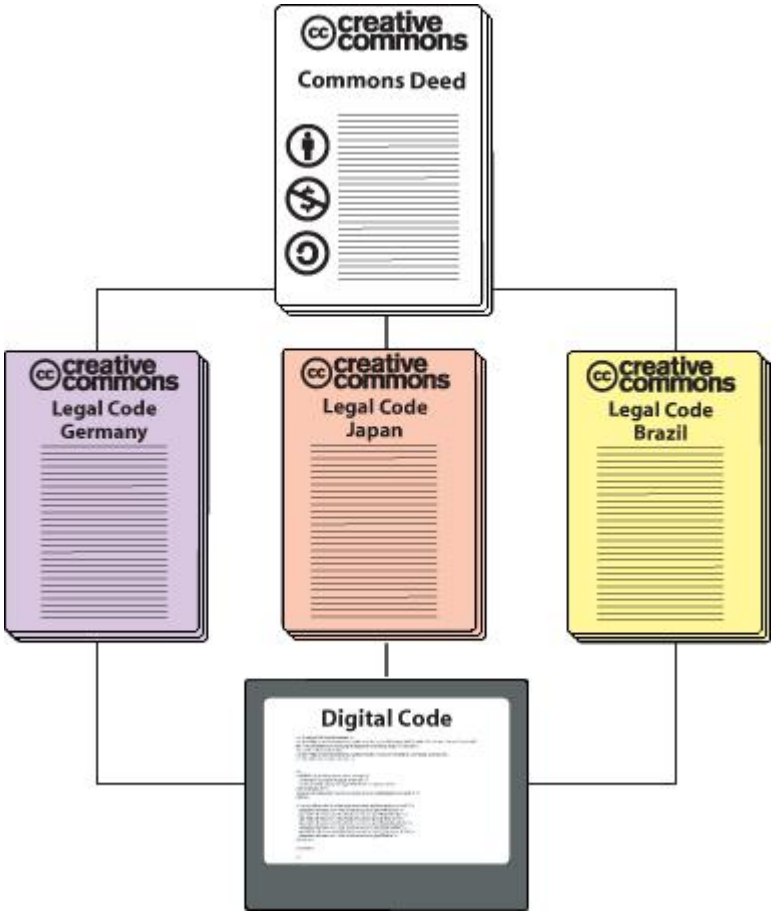

Figure 5.4: The three layers of the Creative Commons Model (from: `http://creativecommons.org/international/`).

4. Attribution Non-Commercial (by-nc) [33]
5. Attribution Non-Commercial Share Alike (by-nc-sa) [34]
6. Attribution Non-Commercial No Derivatives (by-nc-nd) [35]

Attribution is the most accommodating license offered, since the user is free to do whatever he/she wants with the licensed work as long as he gives credit to the licenses/author, while Attribution Share Alike comes closest to open source software licenses. Attribution Non-Commercial Share Alike demands that all derivatives have to be used non-commercially. The most restrictive of Creative Commons licenses is Attribution Non-Commercial No Derivatives: it is also called the "free-advertising license" inasmuch as the licensed worked may be downloaded and shared, but not modified or used commercially.

Creative Commons provides users also the option to let their work fall into the public domain (though this option is not valid in every country outside the US), or to choose any of the GNU copyleft licenses or even the permissive BSD license.

5.3.2 E-Commons and Scientific Research

The Enclosure of the E-Commons

The tragedy of the commons. The patenting of the human genome [36] and the European Database Protection Directive which extends intellectual property rights over mere compilation of facts [36, 33], are in the eyes of James Boyle only two examples for what he calls the "enclosure of the intangible commons of the mind" (a similar expression for the same phenomenon is used by Yochai Benkler [18]. The latter refers to the expansion of intellectual property into the area of uses, works or aspects of works that used to regarded as uncopyrightable. The traditional frontiers of intellectual property rights are under attack [36], questioning the old assumption that the raw materials of scientific research, i.e. ideas, data and fact, should remain in the public domain and not become proprietary [36].

Before we proceed, it is important to note that Boyle obviously uses the term "commons" in its wider meaning, i.e. as both covering the public domain in the strict sense and stretching over into the area of intellectual property (see above).

Now even if the enclosure of the e-commons in some ways parallels the state-promoted transformation of common land into private property in the 19th century [36], there are also dissimilarities between the commons of the mind and its earthy counterpart. Indeed common land is a rival resource inasmuch as many individual uses of the latter mutually exclude each other. Herdsmen who roam the same common pasture compete with each other as to its use and may eventually ruin it: since it is to the immediate benefit of an individual herdsman to add one more cow to his herd, there is no incentive for each one of them to prevent over-grazing of the commons. A "tragedy of the commons" seems to be the outcome: rival resources that are not individually owned inevitably are overexploited [36, 182]. However, such a tragedy does not occur with respect to a commons

[33] http://creativecommons.org/licenses/by-nc/3.0/legalcode
[34] http://creativecommons.org/licenses/by-nc-sa/3.0/legalcode
[35] http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode

that is non-rival - such as in fact the e-commons: there is no limit as to how many times an MP3 is downloaded or a poem is read on the Web [36].

Arguments for and against the sustainability of the e-commons. Defenders of the enclosure of the e-commons therefore prefer to argue that the problem with the informational commons is that there is no incentive to create this resource in the first place. Indeed, information resources are not only non-rival, but also non-excludable: one unit of such a good may satisfy an unlimited number of users at no marginal cost at all [36]. Boyle quite plausibly objects that the Internet compensates this apparent deficiency by also reducing production and distribution costs, while enormously enlarging the market [36]. Moreover, the technologies of the Internet also facilitate quick detection of illegal copying, such that it is not obvious that copyright holders see their privileges diminished through the advent of the Web (ibid.).

Another argument in favour of the enclosure of the e-commons is the growing impact of information-based products in the world economy. However, one may reply that since information products are built out of parts of other information products, and thus every information item constitutes the raw material for further innovation, each additional extension of individual property into the e-common reduces access to and increases the cost of each new product and innovation. Hence, the enclosure of the e-commons may do more harm to innovation that good [36].

As to the question what incentives or motivations there are for building the resources that make up the e-commons - whether it is for prestige, improving one's resume, the satisfaction of exerting one's skills and creativity, or at least partly because of sheer altruistic virtues and values (as claim Benkler and Nissenbaum [19] - it appears be spurious. Indeed, in a global network with a large number of members, there will be always enough talented people that will be willing to contribute to the creation and evaluation of information products, if production and distribution costs are near to zero [36]. Under one condition, however: without centralised supervision, large-scale projects have to be modular in order to allow for an efficient division of labour (ibid.). Open source development is the paradigm of a distributed and non-proprietary creation, a "commons-based peer production" [19], but so has been scientific research and the development of artistic movements long before the existence of the Internet [36].

Distributed creation is also appropriate for capital-intensive projects, at least in the case of science, which more and more relies on data- and processing-intensive models. Lay volunteers have been successfully recruited to the task of distributed data scrutiny, as for example in NASA's "Clickworkers" experiment which recurred to volunteers for the analysis of Mars landing data[36]. Another example for large-scale distributed information production in the field of bioinformatics is the open-source genomics project (www.ensembl.org) [36, 39, 19]. Thus, against economical prejudice in favour of free market competition based on individual property, distributed creativity in an information commons are certainly viable [36].

A false Manichaeism? The danger of the enclosure of the information commons, so Boyle, is that "propertisation is a vicious circle". He argues that in order to achieve optimum price discrimination with proprietary information goods that have no substantial marginal costs, the holders of intellectual property rights will demand ever greater extension of the realm of individual property

[36] `http://clickworkers.arc.nasa.gov/top`

into the information commons [36, 35]. However, the fundamental reason for the tendency to transform the e-commons into private property may be a cognitive bias against openness of systems and networks as well as non-proprietary creation - an aversion which may be due to the fact that our everyday experience of property is that over tangible resources for which the ”tragedy of the commons” indeed holds [32]. Hence the necessity to adapt our conceptions of property to the non-tangible commons of the mind (ibid.).

While one may agree with Boyle’s general concern about the enclosure of the e-commons and deplore the propertisation of the raw material of scientific research, it is appropriate to qualify an excessively Manicheistic view of the dynamics between intellectual property and e-commons. In general, the notions of public good and private ownership are by no means mutually exclusive. A classical example of non-excludable private goods are privately owned lighthouses in 19th century Great-Britain: the service provided by a lighthouse, namely the aid for navigation through the emitted light- or sound signal, cannot be reserved to a few ships [88, 62]. In a sense, e-commons resources are digital-era examples of a possibly private goods that are non-excludable.

It is certainly true that in a first stage reducing the extent of the public domain also means pushing back the frontiers of the e-commons. However, as Boyle himself concedes, the e-commons not only stretches into the area of intellectual property, but actually presupposes intellectual property rights in a crucial respect. As we have seen, copyleft licenses constitute a pillar of the e-commons. But these are copyright licenses that neutralise the monopolistic tendencies inherent in intellectual property. While it is true that intellectual property and public domain correspond to each other like figure and background, and hence each widening of the scope of the former diminishes the extent of the latter, this is not so for the relation between intellectual property and the e-commons. Paradoxically, the extension of copyright means a potential increase of the commons of the mind, provided copyleft licensing keeps up with propertisation. Furthermore, e-commons and intellectual property do not exclude each other in terms of their associated business models: indeed, there is (maybe anecdotal) evidence that some information goods may well be simultaneously available both in the e-commons and on the proprietary marked, without any prejudices to sales in the latter [34]. Not only academic works like Yochai Benkler’s ”The Wealth of Networks” [18] and James Boyle’s ”The Public Domain” [37], but also science fiction novels like ”Down and Out in the Magic Kingdom” by Cory Doctorow have sold considerably well despite being available either in the public domain or under a Creative Commons license (ibid.).

The explanation of this peaceful co-existence may of course reside in the fact that at present, paper copies and electronic copies of a text have complementary uses: pdf-copies are easier to search and quote, while books are more comfortable to carry around or to keep on the bedside table (even more so than print-outs). Of course, the future dissemination of e-book readers may alter this equilibrium. In the case of other media, like music, the comparative advantages of having a hard copy besides the electronic copy may be too marginal to allow for such a harmonious co-existence. For example, the quality of the music as registered on a CD may still be higher than that of an MP3, but for anyone save aficionados of classical music, i.e. for the large majority of consumers that enjoy music as a mere entertainment, it makes no difference to listen to a CD player instead of enjoying the same piece or song on a MP3 player, which has the additional advantage of huge storage capacities.

Self-archiving and Open Access

Authorship vs. Copyright. A distinction which goes often unnoticed is the one between authorship and copyright [126]. Authorship is intellectual priority or "parentship" with respect to an idea or set of ideas, while copyright is the ownership with regard to its expression. Infringement of copyright is theft of text or piracy, while infringement of authorship is theft of ideas or plagiarism. If someone publishes, e.g. reprints, a text without asking the permission of the copyright owner, she may not necessarily also be guilty of plagiarism, which would be the case if she would republish the text under her name. Also, in contrast to copyright, which may be transferred, authorship is unalienable: you can never loose the authorship of your own discoveries and ideas, whereas, in the case of copyright, you can decide to sell or give away the rights on your writings.

The modern conception of scientific authorship was shaped around the birth of the Royal Society and its publication series, The Philosophical Transactions, started in 1665. The community of natural philosophers that founded the Royal Society established some standards and practices related to scientific authorship that are still in force today. For example, they decided that a scientific author cannot "own" his or her own discovery: science writing is a way of reporting about facts of nature, and nature cannot be object of copyright. The members of the Royal Society also introduced an early form of peer-review: new ideas or discoveries were "informally" discussed in the meetings of the Royal Society, and, upon approval by the community of peers, published in the Philosophical Transactions [24].

This distinction between authorship and copyright is especially crucial for scientific literature, which is, in contrast to the majority of the published works, a give-away literature: authors of research papers and books do not seek (and generally do not receive) any royalties, but impact, that is the distribution, recognition and exploitation of their work by their peers. It is on the basis of impact that academics built their career and hence their income [126, 128].

This means that unlike authors of non-give-away works who earn their keep in form of royalties, researchers are less worried about piracy than about plagiarism, i.e. the denial of authorship, since their main concern is that their ideas circulate and gain recognition among their peers. Of course, this does not entail that authors of scientific works would be delighted if their papers and books were pirated; in most cases, they still want to retain control over where their work appears, whether credited or not. However, any obstacle to accessing their works and hence to the impact of their ideas jeopardises their main source of income [126].

Self-archiving and Open Access. Based on this insight, Stevan Harnad, a prominent defender of self-archiving and open access publishing, has been one of the most vocal critics of the traditional subscription-based business model for peer-reviewed scientific journals. Subscription fees have reached a level of about 2000 Euros, which means that research institutions not only in the developing countries have serious difficulties to pay access to refereed journals for their members [126]. In other words, subscription tolls have become access barriers and thus also impact barriers.

Now, peer review is essential for quality assessment and certification of scientific research papers and hence for the academic reputation of their authors; as such it is the only service provided by serious scientific journals in which researchers are really interested [126]. But it has been estimated that the review costs only constitute about 10 percent of the total subscription tolls [126].

The long-term solution advocated by Harnad is the spreading of electronic open-access journals, where publication costs are ideally minimized and are paid by the institutions that host the authors (the so-called "green route"), such that readers can access papers for free [128]. We will return to this topic in the next section.

Meanwhile there is a cheap alternative: self-archiving of pre- and post-prints in institutional eprint archives[37], which has been practised by physicists since 1991. Some publishers, like Springer, provide copyright transfer agreements that explicitly authorize authors to self-archive a personal copy of the refereed and published version, i.e. the so-called post-print, of their paper [126]. In case no such clause can be negotiated, there is a simple and completely legal strategy to circumvent restrictive copyright, namely by self-archiving the preprint and the corrigenda separately (the so-called "Harnad-Oppenheimer strategy": [126, 218]. Of course, this strategy applies only provided the publishers do not request a minimal delay for self-archiving preprints of the final version!

Importing the Open Source philosophy into the world of scientific publishing. By now, the scientific community has realised that Open Access greatly facilitates the circulation of ideas and scientific results, but does not seem to have yet fully grasped its potential for reducing the reckless multiplication of the scientific literature. This issue could at least partially be addressed by introducing the practice of re-use in the production of scientific texts. In an Open Access world, such practice is in principle possible, since the author(s) retain the copyright with respect to their work, while publishing it under a relatively permissive license such as the Creative Commons Attribution license (CC-BY), which explicitly allows for re-use of the text. PLOS[38] and Springer's Open Choice[39] are examples of publishers applying such license. Nonetheless, we are still far away from applying the Open Source philosophy to scientific literature, with the exception of teaching and reference material.

We have seen that free licensing is the conditio sine qua non for "commons-based peer production" [19], i.e. for the distributed creativity that has been the reason why open source software development has been so successful. The license to re-use and modify, together with a peer review in vast global communities, allows for a large-scale incremental optimization of any resource in the e-commons. But while science has applied this model of optimization for the development of ideas and theories, scientific writing is still largely based on the cooperation of small numbers, if not on the romantic cliché of solitary creation.

Online collaborative encyclopaedias such as Wikipedia are examples that large-scale commons-based distributed production can also be harnessed to the creation and improvement of texts. But while this strategy has been applied to technical manuals and teaching material under the GFDL license, especially in Computer Science, even before the Wikibooks initiative[40], this is patently not the case for original research literature.

Currently, the prevailing mentality in academia does not allow for re-use of texts: it is still unthinkable for most academics that you may rewrite a scientific article, correcting some of its flaws (say, a gap in a proof), and publish the new derived version under your name, even if you

[37] see, e.g., `http://www.eprints.org` or `http://hal.archives-ouvertes.fr/`

[38] `http://www.plos.org/journals/license.html`

[39] `http://www.springer.com/open+access/open+choice?SGWID=0-40359-12-161193-0`

[40] `http://en.wikibooks.org`

acknowledge the author of the original paper. Instead you have to write a completely new article that must not substantially textually overlap with the old one. Of course, you may contact the author as the holder of the copyright and negotiate to write a common, improved article. But in even this case, there is an unnecessary waste of time and resources. Conversely, it is still not part of the academic mindset to publish a note for others to re-use and develop, though under a copyleft license, priority of authorship would be safeguarded and derived versions would be iteratively traceable, such that you could in principle gather credits for having sown the seeds of a series of (hopefully) high-quality papers based on your original note.

In the absence of empirical findings on this topic, one may only speculate with more or less plausibility on the reasons why commons-based peer production is not applied to original research literature. First, many academics, especially but not exclusively in the humanities, regard the actual writing of the text not just as a passive registration of ideas, but as contributing to the development of these ideas. That is, in the eyes of many scientists, there may not be a clear-cut separation between the creation of scientific ideas and the production of the texts in which the former are embedded. Since authorship of ideas is not negotiable for researchers since it is career-building, the ideology of the inseparability of ideas and texts, form and content, would partly explain why scientific authors, unlike software designers, do not open up their works for others to modify freely.

Another obstacle to importing the open software philosophy to the production of scientific writings is the fact that the author of the original work may not agree with the ideas expressed in derivative works. Indeed, the author of the original work would need to be included in the list of the author(s) of the derived work, and hence would be attributed responsibility for the ideas expressed in the latter, at least under the current conception of authorship.

Finally, such an innovative way of producing scientific articles would presuppose not changes in the review and crediting process. E.g., drafts intended as seeds of more developed research papers would have to be evaluated differently as fully developed articles.

All three points raised as possible obstacles to applying commons-based peer-production to original research literature, namely the (apparent) inseparability of the creation of ideas and the creation of texts, the problems of attributing responsibility for the derived works in the current authorship model and the necessary transformation of the crediting process point to the necessity of changing academic mentality towards a more collaborative conception of scientific writing and an open model of authorship, close to, but not necessarily identical with, the Open Source philosophy.

5.3.3 Licensing in scientific publishing

The previous section sought to review the whole field of copyright and licensing issues. This section goes in depth within the scientific publishing industry to briefly review, from an historical point of view, the evolution of copyright/licensing issues within the sector. The issues of copyright and licensing in the scientific publishing industry can be regarded as two sides of the same coin, since they are strongly correlated concepts. While an extensive analysis of copyright and licensing was presented before, in this section we briefly review how copyright and licensing might be contextualized within the scientific publishing industry. The section is organized as follows; first, we review the configuration of the market with its main actors and dynamics. Then we review

the evolution of copyright and licensing issues over time. This section is based on a voluminous existing literature about the publishing industry [77, 83, 141, 142, 195, 202, 204, 147, 302, 305, 309].

Copyright and licensing in the publishing industry

The link between copyright and licensing within the scientific publishing industry might be summarized as follows: the intellectual property of the work arranged by a researcher is, by default, owned by the creator itself. The creator might want to the right to use (namely copyright/copyleft) a particular object might be given (namely licensed) to another actor. In the publishing industry the object which is usually subject to copyright is a paper or a monograph. The right to use a paper or monograph is usually an exclusive privilege of the author(s). However, the author(s) may want to license the right to use (i.e. read, disseminate, etc.) the paper to someone else (for instance a publishers or a user). As this simple case shows, there are several ways copyrighted material can be given to someone else. After describing the main actors of the scientific publishing and their relationship, we will assess how copyright and licensing have been declined within the scientific publishing industry before and after the advent of the Internet.

Main actors in the scientific publishing industry. The current scientific publishing industry might be described as having three main classes of actors [26]:

- *Authors.* These are commonly researchers that produce scientific papers or monographs. While in a traditional market the intellectual work of an employee usually belongs to the organization she/he belongs to, this is not the case within the scientific publishing industry. The researcher (as an employee of a University or research institute) retains the intellectual property of her/his work. As researchers build up their careers on the bias of their ability to publish in hi-rated journals, the former usually yield their intellectual property of a particular work to journals. A recent research study estimates the number of papers published in reviewed scientific journals in about 1,300,000 [27].

- *Publishers.* These actors gather intellectual work from authors and manage it to certify its "goodness" and to disseminate it through the publishers' channels. Most of the publishers fall in one of the following two categories. The first category includes commercial/for profit firms while the second includes non-profit organizations. This latter category is made of learned societies and university press. This differentiation is key as in the last decades non-profit organizations have pioneered new licensing models. Currently commercial publishers account for about the 65 per cent of the market share while learned societies' publishers account for 30 per cent. University presses account cover the remaining 5 per cent of the market.

- *Libraries.* These actors buy books and periodicals journals from publishers. Libraries are quite particular actors as they are customers of the scientific publishing industry while they are not those users that read books and journal. Thus, an inconsistency exists between the willingness of single readers to get access to most scientific knowledge the possible and libraries spending availabilities. Currently libraries cover the 65 per cent of the market while

private institutions cover a 30 per cent of the market share. Single readers do not appear as customers as they usually get services for free by libraries.

The concrete relationships between these actors as well as copyright and licensing issues have changed over time. One of the main elements to influence the market has been the Internet. Indeed, Internet has offered new ways to access to scientific knowledge which still are not exhaustively explored and experimented. To fully understand how the Internet has changed copyright and licensing policies and practices we will now review them as they have been before and after the Internet.

Copyright and licensing in scientific publishing before Internet. Before the wide-scale commercial adoption of the Internet, hard copy-based scientific journals represented an essential channel for the diffusion of scientific knowledge.

In this period, publishers could be considered as monopolists of the market as usually the copyright of scientific publications were transferred from authors to them. It is important to note that while authors usually gave away the right to exploit their works, they still retained the scientific authorship (see above). This aspect is very important as authors do not gain from 'selling' their works (monographs or papers); rather their career is based on the ability to publish scientific work with well recognized publishers. As a matter of fact, publishers were working as (1) acquirers of copyrights from authors and as (2) suppliers of licensing policies for the use of copyrighted material. We will review these two main points in detail.

Copyright policy. Traditionally authors transferred their copyright to publishers (either profit or not for profit). It could happen that in particular cases, such as for US government employees, a full copyright transfer was not possible. Usually in this case only a limited part of the transfer was executed. In some other cases, (mostly when contracting materials from companies for professional books) publishers did not obtain the copyright (because the company still wanted to keep this) but exclusive print and distribution rights. *Licensing policy.* Publishers licensed access to their copyrighted material as subscription to journals or by selling books. Libraries were paying the bill. This model has been termed *reader pay* as the right to use the intellectual work is paid by the readers. Usually the commercial strategy of publishers was that of contracting directly with universities and libraries. In particular this strategy allowed differentiating the prices of journals' subscriptions on the basis of each library characteristics.

Copyright and licensing in scientific publishing after the Internet. After the Internet the monopoly of publishers as summarized above started to decrease slowly. Indeed, the new technological opportunities given by the Internet as well as the social and political concerns brought by the *serials crisis*, affected the well established copyright and licensing polices.

Copyright policy. During this period, publishers lost their monopoly in the acquisition of the author's copyright. Indeed, the sharp fall of publication and dissemination costs triggered by the Internet have allowed the birth of different ways to publish intellectual material. In some cases, this opportunity was reflected in the possibility for authors to retain their copyright of their intellectual work. This latter situation was explored by the various Open Access experiments as

outlined below (for a review, please see the website of the Budapest Open Access Initiative[41].

Licensing policy. The birth of Internet brought consistent changes both in terms of how copyright and licensing have been intended and in terms of how these polices have been applied from a commercial point of view.

First, the commercial application of the old ”reader pay” policies has changed. Indeed the almost-zero dissemination costs made possible by the Internet push publishers to sell bundles of journals ”access rights” to libraries. This commercial strategy was called *big deal*, meaning that libraries were offered a set of journal in a single deal. A similar policy is *Core+peripheral* where a small number of core publications are in inserted in the deal. Other subscription policies were tested such as the National License. This latter is a sort of national reading license that covers all the public libraries as well as education and research institutes. Other less known applications are based on a Pay per view (PPV) policy. This latter allows readers paying a fee to get access to a single paper. Among other publishers, this policy has been adopted by journals of the Cambridge University.

Second, initiatives such as the Open Access supported the publication of scientific material while allowing authors to retain the copyright of their intellectual work. The general idea of this approach - that has gained momentum in the last decade - is that scientific knowledge should be completely free and unrestricted access should be granted to scientists. While there are different open access dimensions (see Table 5.1), we will focus on the so-called gold and green routes as they are the most known and applied.

Green Route	The author can self-archive at the time of submission of the publication whether the publication is a grey literature, a peer-reviewed journal publication, a peer-reviewed conference proceedings paper or a monograph
Gold Route	The author or author institution can pay a fee to the publisher at publication time, the publisher thereafter making the material available ”free at the point of access
Preprints	Preprints are articles that are pre-peer-review
Postprints	Postprints are articles that are post-peer-review
Eprints	Eprints can be either preprints or postprints but in electronic form
White Literature	White literature is peer-reviewed, published articles
Grey Literature	Grey literature is preprints or internal ”know-how” material

Table 5.1: Dimensions of Open Access [156]

Open Access Models

The Gold Route. The *gold route* of open access is the commercial version of the open access philosophy. It implies the author or author’s institution pays a fee to the publisher for publishing peer-reviewed research, the publisher thereafter making the material available *free* to all. No subscription to journals is necessary to read an Open Access certified journal or paper.

[41] `http://www.soros.org/openaccess`

The Public Library of Science [193] cites the "Bethesda Statement on Open Access Publishing" when suggesting that an Open Access Publication is one that meets the following two conditions:

- The authors and copyright holders grant to all users a free, irrevocable, worldwide, perpetual right of access to, and a license to copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship, as well as the right to make small numbers of printed copies for their personal use.
- A complete version of the work and all supplemental materials, including a copy of the permission as stated above, in a suitable standard electronic format is deposited immediately upon initial publication in at least one online repository that is supported by an academic institution, scholarly society, government agency, or other well-established organization that seeks to enable open access, unrestricted distribution, interoperability and long-term archiving.

There are three main options of gold open access (OA) publishing licences 5.2:

- Immediate full OA: the entire contents of the journal are made freely available immediately on publication. Well-known examples is the PLoS Biology Journal or the BioMed Central, the biggest OA publisher (owned by Springer Science + Business Media) [268].
- Hybrid and optional OA: here only part of the journal content is made immediately available. There are at least two distinct models:
  - The articles immediately available but requires a subscription to access other "value added" content such as commissioned review articles, journalism, etc.
  - The journal offers authors the option to make their article OA in an otherwise subscription-access journal in return for payment of a fee (e.g. Springer's Open Choice or OUP's Oxford Open schemes). This programs usually leaves the author the opportunity to retain the copyright by scientific publishers maintain the right to commercially exploit the paper.
- Delayed OA: the journal makes its contents freely available after a relatively short period, typically 6-12 months. This model is mostly used by learned societies [27].

Currently, almost all of the worldwide scientific publishers are offering some kind of open access [206].

The Green Route. The *green* route to open access states that authors should be free to self-archive on public open archives their articles. The green route to open access might be divided in institutional archives and repositories 5.2.

Open access archives are typically subject or discipline-based, offering open and free access to pre-print and/or post-print papers in a particular discipline or subject area. On open access

archives, also called subject-based archives, both pre-print articles (articles that have been submitted for publication but not yet accepted) and post-prints (such as articles that have been accepted for publication and/or published) can be published. Usually, they collect documents in a specific discipline and their main objective is to allow for a quicker and more efficient dissemination of papers that are deposited by the authors themselves.

Open access repositories are typically institutionally based, offering the same level of open and free access to the work and outputs of particular institutions (for example, universities or research institutes). Both rely upon authors and/or their institutions posting material to the archive or repository (usually called *self-archiving*). Open access repositories operate in the same way as open access archives, but they are associated with an organization, such as a university or research institute, rather than a subject area or discipline. Currently, several institutions are using the open access repositories to store papers of their affiliates. These initiatives might be divided in two main categories. The first category encompasses those initiatives that are grounded on some sort of law. An example in this sense is the National Library of Medicine managed by the north-American National Institutes of Health (NIH). The policy of this institute imply that all the peer reviewed papers, fully or partially funded by NIH, must be stored, with no exception, in the national repository [50]. The second category encompasses those initiatives that are based on a "voluntaristic basis". An example in this sense is the Harvard Law School policy. This school "suggests" all its members to make their papers available on an online repository. Every following agreement with publishers must take into account that the document is already available on the Web [276].

The House of Commons enquiry concluded that "institutional repositories have the potential greatly to increase the speed, reach and effectiveness of the dissemination of research findings" [142]. The Wellcome Trust noted that [309]:

> ... the existence of a central archive could transform the market. Access to all UK publications would be possible and would act as a brake on excessive pricing". They would benefit authors, readers and institutions: authors would see their articles made available to a wider audience; readers would be able to access articles free of charge over the Internet; and institutions would benefit from having an online platform on which to display their funded research.

The green route to open access is still to be explored [131]. Indeed, it is not really clear how this route will change copyright policies. While within the open access framework, the posting of an already published paper to open archives or repositories does not raise copyright issues, the situation changes when considering a paper which is published within a *subscription based* journal. Here, the usual policy for publishers is to acquire the copyright from authors. Thus the latter could not legally disseminate their work any more.

Self-archiving must be specifically authorized by the publisher. In this sense, it is worth to note that the data provided by the SHERPA partnership[42] shows that an approximate 95 per cent of 523 international publishers allows some kind of self archiving: the 31,7 per cent allows the self archiving of pre-prints only while the 63,2 allows the self-archiving of both pre- and post- prints. However, as mentioned above, it appears that there is another legal way of dealing with copyright issues as to self-archiving of post-prints, namely the "Harnad/Oppenheimer strategy" of archiving

[42] `http://www.sherpa.ac.uk/`

the pre-prints and the corrigenda separately, under the proviso that the copyright transfer for the post-print does not explicitly exclude this procedure [126, 218].

Business models	**Main features**	**Representative Actors**	**Market Share**
Subscription	Publishers acquire copyright	Elsevier, Springer	85-90
Full open access (gold)	Papers immediately available online	Public Library of Science, BioMed Central	~4+6
Delayed open access:	Available after a short period of time	Learned Publishing	<4
Hybrid open access	Open choice or pay per view	Elsevier, Springer	4
Self-archiving (green)	Archives, Institutional repositories	CERN, Wellcome Trust	n.a.

Table 5.2: Different Business and Licensing models

5.4 Conclusion

In this section, we have attempted to gauge the impact of the Social Web both on the production and the distribution of scientific knowledge, focusing on intellectual property rights. After providing an overview on the different innovative features and services of the Web 2.0, we have reviewed the notions of copyright, scientific authorship, and e-commons. The latter is defined as comprising both works that are in the public domain and works that are "copylefted", i.e. licensed by their authors to be copied, shared or modified by anyone provided the copies or modified versions are distributed under the same terms. We have listed various types of free public licenses, both copyleft licenses in the narrow sense and the copyright licenses provided by Creative Commons. We have discussed the relationship between scientific research and the e-commons and have assayed the changes of the copyright and licensing practices in the scientific publishing industry, concluding with overview of the main Open Access models.

This section also outlined the dynamics surrounding the copyright and licensing issues within the scientific publishing industry. As Table 5.2 showed, almost all of the about 23,000 scientific journals still rely on a subscription business models where authors give their copyright to publishers. Nonetheless, the open access approach, with its different versions, is gaining market shares. By rule, this latter approach allows authors to retain the copyright of their articles. Some of the major scientific publishers offer a slight different version of the open access. In some of their subscription based journals, publishers allow the author(s) to have their journal articles made available with full open access in exchange for payment of a basic fee and to retain the copyright. Scientific publishers still maintain the full right to commercially exploit the paper. In more recent years, several funding agencies adopted policies which invite/force authors to publish in open choice journal. Although this latter strategy is quite recent, it is reasonable to think that it will shape the future evolution of the copyright/licensing issues.

Bibliography

[1] A. NIGAM AND N. S. CASWELL. Business artifacts: An approach to operational specification. *IBM Systems Journal 42*, 3 (2003), 428–445.

[2] A.D. GORDON. *Classification, Monographs on Statistics and Applied Probability, Second edition*. Chapman-Hall/CRC, 1999.

[3] ADOMAVICIUS, G. AND TUZHILIN, A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Trans. Know. Data Eng 17*, 6 (2005), 734–749.

[4] ALLISON, D. Design notes for tiny basic. *SIGPLAN Notices 11*, 7 (1976), 25–33.

[5] ALONSO, G. AND CASATI, F. AND KUNO, H. AND MACHIRAJU, V. *Web services. Springer*. Springer, 2003.

[6] ALTHOUSE, B. M., WEST, J. D., BERGSTROM, T. AND BERGSTROM, C. T. Differences in impact factor across fields and over time. *JASIST 60*, 1 (2009).

[7] ALTMAN, A. AND TENNENHOLTZ, M. Ranking systems: The pagerank axioms. In *roceedings of the 6th ACM Conference on Electronic Commerce (EC-05), Vancouver, Canada* (2005), pp. 1–8.

[8] ANITA DE WAARD, G. T. The abcde format enabling semantic conference proceedings. In *SemWiki* (2006).

[9] ANTELMAN K. Do open-access articles have a greater research impact? *College and Research Libraries 65*, 5 (2004), 372–383.

[10] ASHRI, R., RAMCHURN, S. D., SABATER, J., LUCK, M., AND JENNINGS, N. R. Trust evaluation through relationship analysis. In *Proceedings of Fourth International Joint Conference on Autonomous Agents and Multiagent Systems* (2005), pp. 1005–1011.

[11] B. LUDASCHER, I. ALTINTAS, D. H. CHAD BERKLEY, E. JAEGER-FRANK, M. JONES, E. LEE, J. TAO, AND Y. ZHAO. Scientific workflow management and the kepler system. *Concurrency and Computation: Practice and Experience, Special Issue on Scientific Workflows* (2005).

[12] BAADER, F. AND MCGUINNESS, D.L. AND NARDI, D. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[13] BARNES J. Proof and the syllogism. *E. Berti (ed.)* (1981), 17–59.

[14] BARRY WELLMAN AND JANET SALAFF. Computer networks as social networks: Collaborative work, telework, and virtual community. *Annual Review of Sociology 22* (1996).

[15] BAYER A.E., FOLGER J. Some correlates of a citation measure of productivity in science. *Sociology of education* (1966), 381–390.

[16] BENDER, S., AND FISH, A. The transfer of knowledge and the retention of expertise: the continuing need for global assignments. *Journal of Knowledge Management* (2000), 125–137.

[17] BENKLER, Y. Free as air to common use: First amendment constraints on enclosure of the public domain. *New York University Law Review 74* (1999), 354–446.

[18] BENKLER, Y. *The Wealth of Networks*. Yale University Press, 2006.

[19] BENKLER, Y. AND NISSENBAUM, H. Commons-based peer production and virtue. *The Journal of Political Philosophy 14*, 4 (2006), 394–419.

[20] BENOS D.J., BASHARI E., CHAVES J.M., GAGGAR A., KAPOOR N., LAFRANCE M., MANS R., MAYHEW D., MCGOWAN S., POLTER A. AND OTH. The ups and downs of peer review. *Advances in Physiology Education 31*, 2 (2007), 145.

[21] BERRY, D.M, MCCALLION, M. Copyright and copyleft, 2001. Eye magazine. At ttp://www.eyemagazine.com/opinion.php?id=117&oid=290.

[22] BERRY, M. W., DUMAIS, S. T., AND O'BRIEN, G. W. Using linear algebra for intelligent information retrieval. *SIAM Review 37* (1995), 573–595.

[23] BIAGIOLI, M. Priority, originality, and novelty : Construing the new in science, patents, and copyright, 2009. Unpublished paper presented at the Collège de France, Paris, January 15th 2009.

[24] BIAGIOLI, M., GALISON, P. *Scientific Authorship: Credit and Intellectual Property in Science*. Routledge, New York, 2003.

[25] BIRNBAUM H.K. A personal reflection on university research funding. *Physics Today 55*, 3 (2002), 49–53.

[26] BJÖRK, B-C. A model of scientific communication as a global distributed information system. *Information Research 12*, 2 (2007).

[27] BJÖRK, B-C., ROOS, A. AND LAURI, M. Scientific journal publishing: yearly volume and open access availability. *Information Research 14*, 1 (2009).

[28] BLATTNER, M., HUNZIKER, A., AND LAURETIE, T. P. When are recommender systems useful?, 2007. At arXiv:0709.2562.

[29] BOOCH, G., RUMBAUGH, J. AND JACOBSON, I. *The Unified Modeling Language user-guide*. Addison-Wesley, 1997.

[30] BORGMAN, C. L. AND FURNER, J. Scholarly communication and bibliometrics. *Annu. Rev. Info. Sci. Tech. 36* (2002), 3–72.

[31] BORNMANN L.,MUTZ R.,DANIEL H.D. Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology 59*, 5 (2008), 1–8.

[32] BOYLE, J. A closed mind about an open world. Financial Times, August 7, 2006. At http://www.ft.com/cms/s/2/64167124-263d-11db-afa1-0000779e2340.html.

[33] BOYLE, J. Public information wants to be free. Financial Times, February 24 2005. At http://www.ft.com/cms/s/2/cd58c216-8663-11d9-8075-00000e2511c8.html.

[34] BOYLE, J. Text is free, we make our money on volume(s). Financial Times, January 22 2007. At http://www.ft.com/cms/s/b46f5a58-aa2e-11db-83b0-0000779e2340.html.

[35] BOYLE, J. Cruel, mean, or lavish?: Economic analysis, price discrimination and digital intellectual property. *Vanderbilt Law Review 53* (2000).

[36] BOYLE, J. The second eclosure movement and the construction of the public domain. *Law and Contemporary Problems 66* (2003), 33–74.

[37] BOYLE, J. *The Public Domain. Enclosing the Commons of the Mind*. Yale University Press, 2008.

[38] BRAINOV, S. AND SANDHOLM, T. Contracting with uncertain level of trust. In *Proceedings of the First ACM Conference on Electronic Commerce* (1999), pp. 15–21.

[39] BRICKLIN, D. The cornucopia of the commons: How to get volunteer labor. At http://www.bricklin.com/cornucopia.htm.

[40] BRIN, L., AND PAGE, S. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems 30* (1998), 107–117.

[41] BROWMAN, H. I., AND KIRBY, D. S. Quality in science publishing. *Mar. Ecol. Prog. Ser. 270* (2004), 265–287.

[42] BROWMAN, H. I., AND STERGIOU, K. I. The use and misuse of bibliometric indices in evaluating scholarly performance. *Ethics Sci. Environ. Polit. 8*, 1 (2008), 1–107.

[43] BRUNO LATOUR. *Science in Action: How to Follow Scientists and Engineers through Society*. Harvard University Press, Cambridge Mass., USA, 1987.

[44] BRUSILOVSKY, P., KOBSA, A., AND NEJDL, W. *The Adaptive Web (LNCS 4321)*. Springer, Berlin, 2007.

[45] BUSH, V. As we may think. *Atlantic Monthly 176* (1945), 101–108.

[46] BUTTERFIELD, S. Geotagging: One day later. At http://feeds.feedburner.com/ r/Flickrblog/ 3/17553027/geotagging_one_.Html.

[47] CALLAHAM M.L., BAXT W.G., WAECKERLE J.F., WEARS R.L. Reliability of editors' subjective quality ratings of peer reviews of manuscripts. *JAMA 280*, 3 (1998), 229–231.

[48] CAMPBELL, P. Escape from the impact factor. *Ethics Sci. Environ. Polit. 8*, 1 (2008), 5–7.

[49] CARBO, J., MOLINA, J., AND DAVILA, J. . trust management through fuzzy reputation. . 2002, vol. , . *International Journal in Cooperative Information Systems 12* (2002), 35–155.

[50] CARROLL, M. Complying with the national institutes of health ("nih") public access policy: Copyright considerations and options. *A joint sparc/science commons/arl whitepaper* (2008).

[51] CARSTENSEN, P.H. AND SCHMIDT, K. Computer supported cooperative work: new challenges to systems design, 1999. At http://citeseer.ist.psu.edu/carstensen99computer.html.

[52] CASATI, F. AND GIUNCHIGLIA, F. AND MARCHESE, M. Publish and perish: why the current publication and review model is killing research and wasting your money. *ACM Ubiquity* (November 2006).

[53] CASTELFRANCHI, C. AND FALCONE, R. Trust is much more than subjective probability: Mental components and sources of trust. In *Proceedings of the 33rd Hawaii International Conference on System Sciences (on-line edition)* (2000).

[54] CECI S.J., PETERS D.P. Peer review: A study of reliability. *Change 14*, 6 (1982), 44–48.

[55] CHARTIER, R. *Lectures et lecteurs dans la France de l'Ancien Regime*. Seuil, Paris, 1987.

[56] CHEN, P., XIE, H., MASLOV, S., AND REDNER, S. Finding scientific gems with google's pagerank algorithm. *J. Informetrics 1* (2007), 8–15.

[57] CHEUNG, W. W. L. The economics of post-doc publishing. *Ethics Sci. Environ. Polit. 8*, 1 (2008), 41–44.

[58] CHIRITA, P. A. , DAMIAN, A., NEJDL, W., AND SIBERSKI, W. Search strategies for scientific collaboration networks. In *Proceedings of the 2005 ACM workshop on Information Retrieval in Peer-to-Peer Networks* (2005), pp. 33–40.

[59] CHO, M. K., JUSTICE, A. C., WINKER, M. A., BERLIN, J. A., WAECKERLE, J. F., CALLAHAM, M. L., RENNIE, D. Masking author identity in peer review: What factors influence masking success? PEER Investigators. *JAMA 280*, 3 (1998), 243–245.

[60] CHRISTOPHER W. FRASER AND EUGENE W. MYERS. An editor for revision control. *ACM Transactions on Programming Languages and Systems 9*, 2 (1987).

[61] CICCHETTI D.V. Referees, editors, and publication practices: Improving the reliability and usefulness of the peer review system. *Science and Engineering Ethics 3*, 1 (1997), 51–62.

[62] COASE, R. The lighthouse in economics. *Journal of Law and Economics 17* (1974), 357–376.

[63] COMSCORE. Social networking goes global. At http://www.comscore.com/press/release.asp?press=1555.

[64] Copyleft - wikipedia. At http://en.wikipedia.org/wiki/Copyleft.

[65] CUGOLA, G. Tolerating deviations in process support systems via flexible enactment of process models. *IEEE Transactions of Software Engineering 24*, 11 (1998).

[66] D. WODTKE AND G. WEIKUM. A formal foundation for distributed workflow execution based on state charts. In *Proc. Int'l Conf. on Database Theory* (1997).

[67] DALTON, R. Peers under pressure. *Nature 413* (2001), 102–104.

[68] DAVID LORGE PARNAS. Stop the numbers game. *Communications of the ACM 50*, 11 (2007).

[69] DAVIS, P. M. AND FROMERTH, M. J. Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics 71* (2007), 203–215.

[70] DAYTON, A. I. Beyond open access: Open discourse, the next great equalizer. *Retrovirol 3* (2006), 55.

[71] DE KERCHOVE, C. AND VAN DOOREN, P. Iterative filtering for a dynamical reputation system. *arXiv: 0711.3964* (2007).

[72] DELLAROCAS, C. *The Digitalization of Word-Of-Mouth: Promise and Challenges of Online Reputation Mechanisms*. INFORMS, 2003, p. En línea.

[73] DIMITRIOS GEORGAKOPOULOS, MARK F. HORNICK, AMIT P. SHETH. An overview of workflow management: From process modeling to workflow automation infrastructure. *Distributed and Parallel Databases 3*, 2 (1995), 119–153.

[74] DONOVAN B. The truth about peer review. *Learned Publishing 11*, 3 (1998), 179–184.

[75] DOURISH, P.; BELLOTTI, V. Awareness and coordination in shared workspaces. In *ACM conference on Computer-supported cooperative work* (1992), pp. 107–114.

[76] DOURISH P, EDWARDS K, LAMARCA A, LAMPING J, PETERSEN K, SALISBURY M, TERRY D AND THORNTON J. Extending document management systems with user-specific active properties. *ACM Trans Info Sys 18*, 2 (2000), 140–170.

[77] ECDGR. Study on the economic and technical evolution of the scientific publication markets in europe. european commission directorate general for research, 2006. At http://ec.europa.eu/research/science-society/pdf/scientific-publication-study_en.pdf.

[78] EDITORIAL. Citation data: the wrong impact? *Nature Neuroscience 1*, 8 (1998), 641–642.

[79] EDITORIAL. Coping with peer rejection. *Nature 425* (2003), 645.

[80] EDITORIAL. Deciphering impact factors. *Nature Neuroscience 6*, 8 (2003), 783.

[81] EDITORIAL. Revolutionizing peer review? *Nature Neuroscience 8* (2005), 397.

[82] Emacs general public license. At http://www.free-soft.org/gpl_history/emacs_gpl.html.

[83] ENSERINK, M. European union steps back from open-access leap. *Science 315* (2007), 1065.

[84] EUROPEAN COMMISSION. Study on the economic and technical evolution of the scientific publication market in europe, 2006.

[85] FABIO CASATI, STEFANO CERI, BARBARA PERNICI, GIUSEPPE POZZI. Workflow evolution. In *ER* (1996), pp. 438–455.

[86] FABIO CASATI, STEFANO CERI, BARBARA PERNICI, GIUSEPPE POZZI. Workflow evolution. *Data Knowl. Eng. 24*, 3 (1998), 211–238.

[87] FAUSTO GIUNCHIGLIA, PAVEL SHVAIKO, MIKALAI YATSKEVICH. Semantic matching. *Encyclopedia of Database Systems* (2009).

[88] FOLDVARY, F. The lighthouse as a private-sector collective good. the independent institute working paper 46. At http://www.independent.org/publications/working_papers/article.asp?id=757.

[89] FRANZEN, M., RÖDDER, S., AND WEINGART, P. Fraud: Causes and culprits as perceived by science and the media. *EMBO Reports 8*, 1 (2007), 3–7.

[90] FREDRIK ARVIDSSON AND ANNIKA FLYCHT-ERIKSSON. Ontologies i. At http://www.ida.liu.se/ janma/SemWeb/Slides/ontologies1.pdf.

[91] GADD E., OPPENHEIM C., PROBETS S. The intellectual property rights issues facing self-archiving. *D-Lib Magazine 9*, 9 (2003).

[92] GARFIELD, E. Citation indexes for science: a new dimension in documentation through association of ideas. *Science 122* (1955), 108–111.

[93] GARFIELD, E. Citation analysis as a tool in journal evaluation. *Science 178* (1972), 471–479.

[94] GARFIELD, E. The history and meaning of the journal impact factor. *JAMA 295*, 1 (2006), 90–93.

[95] GARFIELD, E. AND SHER, I. H. New factors in the evaluation of scientific literature through citation indexing. *Amer. Doc. 14*, 3 (1963), 195–201.

[96] GARFIELD E.,WELLJAMS-DOROF A. Of nobel class: A citation perspective on high impact research authors. *Theoretical Medicine and Bioethics 13*, 2 (1992), 117–135.

[97] GARFUNKEL J.M., ULSHEN M.H., HAMRICK H.J., LAWSON E.E. Effect of institutional prestige on reviewers' recommendations and editorial decisions. *JAMA 272*, 2 (1994), 137–138.

[98] GELLER, P. E. *International Copyright Law and Practice*. Matthew Bender, 2003.

[99] GEMMELL, J., BELL, G., LUEDER, R., DRUCKER, S. AND WONG, C. Mylifebits: Fulfilling the memex vision. In *Proceedings of ACM Multimedia* (2002), pp. 235–238.

[100] GILBERT, J. R., WILLIAMS, E. S., AND LUNDBERG, G. D. Is there gender bias in jama's peer review process? *JAMA 272* (1994), 139–142.

[101] GILES, J. Internet encyclopaedias go head to head. *Nature 438* (2005), 900–901.

[102] GILES, J. Journals submit to scrutiny of their peer-review process. *Nature 439*, 252 (2006).

[103] GINSPARG, P. First steps towards electronic research communication. *Comput. Phys. 8* (1994), 390–396.

[104] GIUNCHIGLIA, F., MARCHESE, M., AND ZAIHRAYEU, I. *Encoding Classifications into Lightweight Ontologies*. Springer, 2007, pp. 57–81.

[105] GIUNCHIGLIA, F., AND ZAIHRAYEU, I. *Lightweight Ontologies*. Springer Verlag, 2008.

[106] GIUNCHIGLIA F., SHVAIKO P., YATSKEVICH M. Semantic schema matching. In *CoopIS, DOA, and ODBASE, OTM Confederated International Conferences, Agia Napa, Cyprus* (2005), pp. 347–365.

[107] GIUNCHIGLIA F., SHVAIKO P., YATSKEVICH M. Discovering missing background knowledge in onology matching. In *17th European Conference on Artificial Intelligance - ECAI 2006* (2006), pp. 382–386.

[108] GIUNCHIGLIA F., YATSKEVICH M. Element level semantic matching. In *Meaning Coordination and Negotiation workshop at ISWC'04* (2004).

[109] GIUNCHIGLIA F., YATSKEVICH M., GIUNCHIGLIA E. Efficient semantic matching. In *Proceedings of the 2nd European semantic web conference (ESWC'05)* (2005), pp. 272–289.

[110] GIUNCHIGLIA F., YATSKEVICH M., SHVAIKO P. Semantic matching: Algorithms and implementation. *Journal on Data Semantics 2*, 9 (2007), 1–38.

[111] GLICKMAN, M. E. Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics 48* (1999), 377–394.

[112] GODLEE, F. Making reviewers visible: Openness, accountability, and credit. *JAMA 287* (2002), 2762–2765.

[113] GODLEE F., GALE C.R., MARTYN C.N. Effect on the quality of peer review of blinding reviewers and asking them to sign their reports a randomized controlled trial. *JAMA 280*, 3 (1998), 237–240.

[114] GOLBECK, J. AND HENDLER, J. Inferring trust relationships in web-based social networks. *ACM Transactions on Internet Technology* (2006), 497–529.

[115] GOLDBERG, K., ROEDER, T., GUPTA, D., AND PERKINS, C. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval 4* (2001), 133–151.

[116] GONG CHENG AND YUZHONG QU. Integrating lightweight reasoning into class-based query refinement for object search. In *Proceedings of the 3rd Asian Semantic Web Conference* (2008), pp. 449–463.

[117] GONG CHENG, WEIYI GE, YUZHONG QU. Falcons: Searching and browsing entities on the semantic web. In *Proceedings of the 17th International Conference on World Wide Web* (2008), pp. 1101–1102.

[118] GOODMAN S.N., BERLIN J., FLETCHER S.W., FLETCHER R.H. Manuscript quality before and after peer review and editing at annals of internal medicine. *Annals of Internal Medicine 121*, 1 (1994), 11–21.

[119] GREENBERG, S. AND MARWOOD, D. Real time groupware as a distributed system: concurrency control and its effect on the interface. In *ACM conference on Computer supported cooperative work* (1994), pp. 207–217.

[120] GROZA, T., HANDSCHUH, S., MÖLLER, K., AND DECKER, S. Salt - semantically annotated latex for scientific publications. In *The Semantic Web: Research and Applications, 4th European Semantic Web Conference, ESWC 2007* (2007), pp. 518–532.

[121] GRUDIN, J. Computer-supported cooperative work: Its history and participation. *Computer 27*, 4 (1994), 19–26.

[122] GYÖNGYI AND H. GARCIA-MOLINA AND J. PEDERSEN. Combating web spam with trustrank. In *Proceedings of the International Conference on Very Large Data Bases* (2004).

[123] HAJJEM C., GINGRAS Y., BRODY T., CARR L., HARNAD S. Open access to research increases citation impact. Tech. rep., Universite' du Que'bec a' Montre'al, 2005.

[124] HANANI, U., SHAPIRA, B., AND SHOVAL, P. Information filtering: overview of issues, research and systems. *User Modeling and User-Adapted Interaction 11* (2001), 203–259.

[125] HARNAD, S. Scholarly skywriting and the prepublication continuum of scientific enquiry. *Psychol. Sci. 1* (1990), 342–344.

[126] HARNAD, S. The self-archiving initiative. *Nature 410* (2001), 1024–1025.

[127] HARNAD, S. *Skyreading and Skywriting for Researchers: A Post-Gutenberg Anomaly and How to Resolve it*. Palgrave Mac Millian, 2001.

[128] HARNAD, S. The postgutenberg open access journal. In *The Future of the Academic Journal*. Chandos, 2008.

[129] HARNAD S., BRODY T. Comparing the impact of open access (oa) vs. non-oa articles in the same journals. *D-lib Magazine 10*, 6 (2004).

[130] HARNAD S., BRODY T., VALLIÈRES F., CARR L., HITCHCOCK S., GINGRAS Y., OPPENHEIM C., HAJJEM C., HILF E.R. The access/impact problem and the green and gold roads to open access: An update. *Serials review 34*, 1 (2008), 36–40.

[131] HARNAD, S., BRODY, T., VALLIERES, F., CARR, L., HITCHCOCK, S., GINGRAS, Y., OPPENHEIM, C., STAMERJOHANNS, H. AND HILF, E. The access/impact problem and the green and gold roads to open access. *Serials Review 30*, 4 (2004), 310–314.

[132] HAUPTMAN R. *Ethics and Librarianship*. McFarland & Company, 2002.

[133] HAYTHORNTHWAITE, C. Social networks and internet connectivity effects. *Information, Communication, and Society 8*, 2 (2005), 125–147.

[134] HEATH, T., MOTTA, E., AND PETRE, M. Computing word-of-mouth trust relationships in social networks from semantic web and web2.0 data sources. In *Workshop on Bridging the Gap between Semantic Web and Web 2.0 at 4th European Semantic Web Conference, Innsburck, Austria* (2007).

[135] HERLOCKER, J. L., KONSTAN, J. A., TERVEEN, L. G., AND RIEDL, J. T. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst. 22*, 1 (2004), 5–53.

[136] HIRSCH, J. E. An index to quantify an individual's scientific research output. *PNAS 102*, 46 (2005), 16569–16572.

[137] HIRSCH, J. E. Does the h-index have predictive power? *PNAS 104*, 49 (2007), 19193–19198.

[138] HITCHCOCK S., WOUKEU A., BRODY T., CARR L., HALL W., HARNAD S. Evaluating Citebase, an open access web-based citation-ranked search and impact discovery service. Tech. rep., University of Southampton, 2003.

[139] HOFFMANN, T. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems 22* (2004), 89–115.

[140] HOLMES A. Electronic publishing in science: reality check. *Canadian Journal of Communication 22*, 3 (1997).

[141] HOUGHTON, J. Digital broadband content: Scientific publishing. directorate for science, technology and industry, OECD. Tech. rep., Organisation for Economic Co-operation and Development, 2005. At http://www.oecd.org/dataoecd/42/12/35393145.pdf.

[142] HOUSE OF COMMONS. Scientific publications: Free for all?. the science and technology committee. At http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/39902.htm.

[143] HURST, M. , SIEGLER, M., GLANCE, N. On estimating the geographic distribution of social media. In *ICSWSM* (2007).

[144] HUYNH, T. D., JENNINGS, N. R., AND SHADBOLT, N. Fire: an integrated trust and reputation model for open multi-agent systems. In *Proceedings of the 16th European Conference on Artificial Intelligence, Valencia, Spain* (2004).

[145] INDRAJIT RAYA, JUNXING ZHANGB. Towards a new standard for allowing concurrency and ensuring consistency in revision control systems. *Computer Standards and Interfaces 29*, 3 (2007).

[146] INGELFINGER, F. J. Peer review in biomedical publication. *Amer. J. Med. 56* (1974), 686–692.

[147] INTERNATIONAL ASSOCIATION OF STM PUBLISHERS. An overview of scientific, technical and medical publishing and the value it adds to research outputs. position paper on scientific, technical and medical publishing, April 2008. At http://www.stm-assoc.org/documents-statements-public-co/2008-04_Overview_of_STM_Publishing_Value_Research.pdf.

[148] J. WANG AND A. KUMAR. *A framework for document-driven workflow systems*. Springer Berlin, 2005, pp. 285–301.

[149] JACKSON, A. From preprints to e-prints: the rise of electronic preprint servers in mathematics. *Not. Amer. Math. Soc. 49* (2002), 23–32.

[150] Guarding the guardians: research on editorial peer review. selected proceedings from the first international congress on peer review in biomedical publication. *JAMA 263* (1990), 1317–1441.

[151] Proceedings of the second international conference on peer review in biomedical publication. *JAMA 272* (1994), 91–173.

[152] Proceedings of the third international conference on peer review in biomedical publication. *JAMA 280* (1998), 207–306.

[153] Proceedings of the fourth international conference on peer review in biomedical publication. *JAMA 287* (2002), 2749–2989.

[154] JEFFERSON, T., ALDERSON, P., WAGER, E., AND DAVIDOFF, F. Effects of editorial peer review: a systematic review. *JAMA 287*, 21 (2002), 2784–2786.

[155] JEFFERSON T., WAGER E., DAVIDOFF F. Measuring the quality of editorial peer review. *JAMA 287*, 21 (2002), 2786–2790.

[156] JEFFERY, G. Open access: An introduction. (on line) ercim news edition. At `http://www.ercim.org/publication/Ercim_News/enw64/jeffrey.html`.

[157] JUSTICE, A. C., CHO, M. K., WINKER, M. A., BERLIN, J. A., RENNIE, D., AND PEER INVESTIGATORS. Does masking author identity improve peer review quality? a randomized controlled trial. *JAMA 280*, 3 (1998), 240–242.

[158] KARL FOGEL. Producing open source software: How to run a successful free software project, 2007.

[159] KASSIRER J.P.,CAMPION E.W. Peer review: Crude and understudied, but indispensable. *Journal of American Medical Association 272*, 2 (1994), 96–97.

[160] KATHLEEN WETS AND DAVE WEEDON AND JAN VELTEROP. Post-publication filtering and evaluation: Faculty of 1000. *Learned Publishing 16* (2003).

[161] KATZ, D. S., PROTO, A. V., AND OLMSTED, W. W. Incidence and nature of unblinding by authors: our experience at two radiology journals with double-blinded peer review policies. *Amer. J. Roentgenol. 179* (2002), 1415–1417.

[162] KELLY C.D.,JENNIONS M.D. The h index and career assessment by numbers. *Trends in Ecology & Evolution 21*, 4 (2006), 167–170.

[163] KENT BECK. . embracing change with extreme programming. , vol. 32, no. 10. 1999. *Computer 32*, 10 (1999).

[164] KLEINBERG, J. Authoritative sources in a hyperlinked environment. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, New York* (1998), pp. 668–677.

[165] KLEINBERG, J. M. Authoritative sources in a hyperlinked environment. *Journal of the ACM 46* (1999), 604–632.

[166] KOEPSELL, D. R. *The Ontology of Cyberspace.* Chicago: Open Court, La Salle, IL USA, 2000.

[167] KOOMIN, E., LIPMAN, D., DIGNON, R., AND LANDWEBER, L. Reviving a culture of scientific debate. *Nature Web Focus* (2006).

[168] KOOMIN, E. V., LANDWEBER, L. F., AND LIPMAN, D. J. A community experiment with fully open and published peer review. *Biol. Direct 1*, 1 (2006).

[169] KORNFELD, WILLIAM A. AND HEWITT, CARL. The scientific community metaphor. *IEEE Transactions on Systems, Man, and Cybernetics 11*, 1 (1981).

[170] KRONICK, D. A. Peer review in 18th-century scientific journalism. *JAMA 263* (1990), 1321–1322.

[171] KURTZ, M. J., EICHHORN, G., ACCOMAZZI, A., GRANT, C., DEMLEITNER, M., HENNEKEN, E., AND MURRAY, S. S. Does anyone know the title of this article? *Inform. Process. Manag. 41* (2005), 1395–1402.

[172] KURTZ M.J. Restrictive access policies cut readership of electronic research journal articles by a factor of two. *Harvard-Smithsonian Centre for Astrophysics, Cambridge, MA* (2004).

[173] LAMARCA, A. AND EDWARDS, K. AND DOURISH, P. AND LAMPING, J. AND SMITH, I. AND THORNTON, J. Taking the work out of workflow: Mechanisms for document-centric collaboration. In *Proceedings of the 6th European Conference on Computer-Supported Cooperative Work (ECSCW '99, Copenhagen, Denmark, Sept. 12-16)* (1999), Kluwer Academic, Dordrecht, Netherlands.

[174] LANGE, D. Recognizing the public domain. *Law and Contemporary Problems 44* (1981), 147.

[175] LAURETI, P., MORET, L., ZHANG, Y.-C., AND YU, Y.-K. Information filtering via iterative refinement. *Europhysics Letters 75*, 6 (2006), 1006–1012.

[176] LAWRENCE, P. A. The mismeasurement of science. *Curr. Biol. 17*, 15 (2007), R583–R585.

[177] LAWRENCE P.A. The politics of publication. *Nature 422*, 6929 (2003), 259–261.

[178] LAWRENCE P.A. Lost in publication: how measurement harms science. *Ethics Sci Environ Polit 8* (2008), 9–11.

[179] LEHMANN, S., JACKSON, A. D., AND LAUTRUP, B. E. Measures for measures. *Nature 444* (2006), 1003–1004.

[180] LERNER E. Fraud shows peer-review flaws. *The Industrial Physicist 8*, 2 (2002).

[181] LERNER, J. 150 years of patent office practice, 1999.

[182] LESSIG, L. *The Future of Ideas*. Random House, 2001.

[183] LESSIG, L. The architecture of innovation. *Duke Law Journal 51* (2002), 1783.

[184] LI DING, RONG PAN, TIM FININ, ANUPAM JOSHI, YUN PENG, AND PRANAM KOLARI. Finding and ranking knowledge on the semantic web. In *Proceedings of the 4th International Semantic Web Conference* (2005).

[185] LINDBERG, S. W., PATTERSON L. R. *The Nature of Copyright: A Law of Users' Rights*. University of Georgia Press, 1991.

[186] LINDEN, G., SMITH, B., AND YORK, J. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing 7* (2003), 76–80.

[187] LINK A.M. Us and non-us submissions an analysis of reviewer bias, 1998.

[188] LITMAN, J. The public domain. *Emory Law Journal 39* (1990).

[189] LLOYD, M. E. Gender factors in reviewer recommendations for manuscript publication. *J. Appl. Behav. Anal. 23*, 4 (1990), 539–543.

[190] M. REICHERT AND P. DADAM. Adeptflex: Supporting dynamic changes of workflow without loosing control. *Journal of Intelligent Information Systems 10*, 2 (1998), 93–129.

[191] M. WESKE, G. VOSSEN, C.B. MEDEIROS. Scientific workflow management: Wasa architecture and applications. *Fachbericht Angewandte Mathematik und Informatik 3*, 1 (1996).

[192] MACCALLUM, C. J. ONE for all: the next step for PLoS. *PLoS Biology 4*, 11 (2006).

[193] MACCALLUM, C. When is open access not open access? *PLoS Biol 5*, 10 (2007).

[194] MACCALLUM C.J., PARTHASARATHY H. Open access increases citation rate. *PLoS Biology 4*, 5 (2006).

[195] MADRAS, G. Scientific publishing: Rising cost of monopolies. *Current Science 95*, 2 (2008), 163.

[196] Markup language – wikipedia. At http://www.w3.org/MarkUp/.

[197] MARRIS, E. Should journals police scientific fraud? *Nature 439* (2006), 520–521.

[198] MCNUTT, R. A., EVANS, A. T., FLETCHER, R. H., AND FLETCHER, S. W. The effects of blinding on the quality of peer review: a randomized trial. *JAMA 263*, 10 (1990), 1371–1376.

[199] MEDELYAN, O., LEGG, C., MILNE, D., AND WITTEN, I. H. Mining meaning from wikipedia. *CoRR abs/0809.4530* (2008).

[200] Metadata- wikipedia:. At http://en.wikipedia.org/wiki/Metadata.

[201] MICHAL JACOVI, VLADIMIR SOROKA. The chasms of cscw: a citation graph analysis of the cscw conference. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work* (2006). At http://portal.acm.org/citation.cfm?id=1180875.1180920.

[202] MILMO, D. Publishers watch in fear as a new world comes into view. The Guardian, April 19, 2006. At http://www.guardian.co.uk/technology/2006/apr/19/news.science1.

[203] MIZZARO, S. Quality control in scholarly publishing: a new proposal. *JASIST 54*, 11 (2003), 989–1005.

[204] MORGAN STANLEY. Scientific publishing: Knowledge is power. industry overview report. September 30, 2002.

[205] MORRIS S. The true costs of scholarly journal publishing. *Learned Publishing 18*, 2 (2005), 115–126.

[206] MORRIS, S. Mapping the journal publishing landscape: how much do we know? *Learned Publishing 20*, 3 (2007), 299–310.

[207] MUI, L. Computational models of trust and reputation: Agents, evolutionary games, and social networks. Tech. Rep. PhD thesis, Massachusetts Institute of Technology (MIT), 2002.

[208] NAAK, A., HAGE, H., AND A, E. Papyres: A research paper management system. *E-Commerce Technology and Enterprise Computing, E-Commerce and E-Services, IEEE Conference and Fifth IEEE Conference 0* (2008), 201–208.

[209] NARDI, B.A. AND WHITTAKER, S. AND BRADNER, E. Interaction and outeraction: instant messaging in action. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (2000), ACM Press New York, NY, USA, pp. 79–88.

[210] NEWMAN, M. E. J. Complex networks. *SIAM Review 45*, 2 (2003), 167–256.

[211] NEWMAN, M. E. J. The first-mover aadvantage in scientific publication. arXiv: 0809.0522. 2008.

[212] NEWMAN, M. E. J. Who is the best connected scientist? a study of scientific coauthorship networks. *Physical Review E. 64*, 016132 (2001).

[213] NOYONS E.C.M.,MOED H.F., LUWEL M. Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. *Journal of the American Society for Information Science 50*, 2 (1999).

[214] NYLENNA M., RIIS P., KARLSSON Y. Multiple blinded reviews of the same two manuscripts. effects of referee characteristics and publication language. *JAMA 272*, 2 (1994), 149–151.

[215] ODLYZKO A. The rapid evolution of scholarly communication. *Learned Publishing 15*, 1 (2002), 7–19.

[216] ODLYZKO, A. M. Tragic loss or good riddance? the impending demise of traditional scholarly journals. *Int. J. Hum.-Comput. St. 42* (1995), 71–122.

[217] Ontology. At http://en.wikipedia.org/wiki/Ontology.

[218] OPPENHEIM, C. *The legal and regulatory environment for electronic information*. Infonortics, 2001.

[219] ORR, R. H. AND KASSAB, J. Peer group judgements on scientific merit: Editorial refereeing. Presentation to the Congress of the International Federation for Documentation, Washington DC. 1965.

[220] P. ROTHWELL, C. M. Reproducibility of peer review in clinical neuroscience. *Brain 123* (2000), 9.

[221] PAGE G.,CAMPBELL R.,MEADOWS A.J. Journal publishing in cambridge. *University Press* (1997).

[222] PAZZANI, M. J., AND BILLSUS, D. Content-based recommendation systems. *LNCS 4321* (2007), 325–321.

[223] PERUGINI, S., GONÇALVES, M. A., AND FOX, E. A. Recommender systems research: a connection-centric survey. *J. Intell. Inf. Sys. 23*, 2 (2004), 107–143.

[224] PETER DADAM, MANFRED REICHERT, STEFANIE RINDERLE, MARTIN JURISCH, HILMAR ACKER, KEVIN GÖSER, ULRICH KREHER, MARKUS LAUER. Towards truly flexible and adaptive process-aware information systems. In *UNISCON* (2008), pp. 72–83.

[225] PEW. Pew internet (2007). At http://www.pewinternet.org/pdfs/PIP_Tagging.pdf.

[226] PEW. Pew internet and american life online video 2007. At http://www.pewinternet.org.

[227] PRINGLE J. Thomson scientific finds new opportunities in open access. *KnowledgeLink Newsletter from Thomson Scientific* (2004).

[228] PUJOL, J.M., SANGÜESA, R., AND DELGADO, J. Extracting reputation in multi agent system by means of social network topology. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems, Bologna, Italy* (2002), pp. 467–474.

[229] PURCELL, G. P., DONOVAN, S. L., AND DAVIDOFF, F. Changes to manuscripts during the editorial process: Characterizing the evolution of a clinical paper. *JAMA 280* (1998), 227–228.

[230] R. GUHA, ROB MCCOOL, ERIC MILLER. Semantic search. In *WWW2003* (2003).

[231] R KRISHNAN, L MUNAGA, K KARLAPALEM. Xdoc-wfms: A framework for document centric workflow management system.

[232] RAMCHURN, S. D., HUYNH, T. D., AND JENNINGS, N. R. Trust in multiagent systems. *The Knowledge Engineering Review 19*, 1 (2004), 1–25.

[233] RAVI KUMA AND JASMINE NOVAK AND ANDREW TOMKINS. Structure and evolution of online social networks. In *International Conference on Knowledge Discovery and Data Mining* (2006).

[234] RAY J.G. Judging the judges: the role of journal editors, 2002.

[235] RAYMOND, E. S. *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. O'Reilly, Sebastopol, CA, 1999.

[236] Rdf specification development. At http://www.w3.org/RDF/.

[237] Rdf- wikipedia. At http://en.wikipedia.org/.

[238] RENNIE, D. Fourth international congress on peer review in biomedical publication. *JAMA 287* (2002), 2759–2760.

[239] RICAURTE G.A., YUAN J., HATZIDIMITRIOU G., BRANDEN J., MCCANN U.D. Retraction, 2003.

[240] RICHARD HULL. Artifact-centric business process models: Brief survey of research results and challenges. In *Proceedings of the On The Move Federated Conference* (2008).

[241] RIEGE, A. Three-dozen knowledge-sharing barriers managers must consider. *Journal of Knowledge Management 9*, 3 (2005), 18 – 35.

[242] RINDERLE, S., KREHER, U., LAUER, M., DADAM, P., REICHERT, M. On representing instance changes in adaptive process management systems. In *15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises* (2006), pp. 297–304.

[243] ROBERTS L.W., COVERDALE J., EDENHARDER K., LOUIE A. How to review a manuscript: A down-to-earth approach, 2004.

[244] S. A. WHITE. *Process Modeling Notations and Workflow Patterns*. Future Strategies Inc., 2004.

[245] SABATER, J. AND SIERRA, C. Regret: a reputation model for gregarious societies. In *Proceedings of the Fourth Workshop on Deception, Fraud and Trust in Agent Societies, Montreal, Canada* (2001), pp. 61–69.

[246] SABATER, J. AND SIERRA, C. Reputation and social network analysis in multiagent systems. In *Proceedings of First International Joint Conference on Autonomous Agents and Multiagent Systems* (2002), pp. 475–482.

[247] SABATER, J. AND SIERRA, C. Review on computational trust and reputation models. *Artificial Intelligence Review 24* (2005), 33–60.

[248] SABATER, J., PAOLUCCI, M., AND CONTE, R. Repage: Reputation and image among limited autonomous partners. *Journal of Artificial Societies and Social Simulation 9*, 2 (2006).

[249] SANGER, L. Why wikipedia must jettison its anti-elitism. Kuro5hin. 2004. At http://kuro5hin.org.

[250] SCHAFER, J. B., KONSTAN, J. A., AND RIEDL, J. E-commerce recommendation applications. *ata Mining and Knowledge Discovery 5* (2001), 115–153.

[251] SCHROTER S., BLACK N., EVANS S., CARPENTER J., GODLEE F., SMITH R. Effects of training on quality of peer review: randomised controlled trial. *British Medical Journal 328*, 7441 (2004), 673.

[252] SCHUHMANN, R. Editorial: Peer review per physical review. *Phys. Rev. Lett. 100*, 050001 (2008).

[253] SCOTT CHRISTLEY AND GREG MADEY. Collection of activity data for sourceforge projects. Tech. Rep. TR-2005-15, University of Notre Dame, 2005.

[254] SEGLEN, P. O. Why the impact factor of journals should not be used for evaluating research. *BMJ 314* (1997), 497.

[255] Semantic search – wikipedia. At http://en.wikipedia.org/.

[256] Semantic search survey. At http://swuiwiki.webscience.org/index.php/Semantic_Search_Survey.

[257] Semantic web tools. At http://esw.w3.org/topic/SemanticWebTools.

[258] SERENO, B., SHUM, S. B., AND MOTTA, E. Formalization, user strategy and interaction design: Users' behaviour with discourse tagging semantics. In *Proceedings Workshop on Social and Collaborative Construction of Structured Knowledge, 16th International World Wide Web Conference (WWW 2007)* (2007).

[259] SHARDANAND, U. AND MAES, P. Social information filtering: Algorithms for automating word of mouth. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (1995), pp. 210–217.

[260] SHUTTLEWORTH, M. Talk at DebConf'05.

[261] SIERRA, C., AND DEBENHAM, J. Information-based reputation. In *First International Conference on Reputation: Theory and Technology (ICORE 2009)* (Gargonza, Italy, 2009), M. Paolucci, Ed., pp. 5–19.

[262] SIERRA, C. AND DEBENHAM, J. An information-based model for trust. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-05), Utrecht, Netherlands* (005).

[263] SIMON BUCKINGHAM SHUM, E. M., AND DOMINGUE, J. Modelling and visualizing perspectives in internet digital libraries. In *Proceedings of ECDL'99: Third European Conference on Research and Advanced Technology for Digital Libraries* (1999).

[264] SINGH, M., AND VOUK, M. A. Scientific workflows: Scientific computing meets transactional workflow. In *Proceedings of the NSF Workshop on Workflow and Process Automation in Information Systems: State-of-the-Art and Future Directions* (1996), p. Online Proceedings.

[265] SMIGEL, E. O. AND ROSS, H. L. Factors in the editorial decision. *Amer. Sociol. 5* (1970), 19–21.

[266] SMITH L.C. Citation analysis, 30 lib. *Trends 83* (1981), 85.

[267] SMITH R. Peer review: a flawed process at the heart of science and journals. *JRSM 99*, 4 (2006), 178.

[268] SMITH, R. A great day for science, never mind the bank chaos - this week's boost for open access research could be more important in the long run. *The Guardian 14*, 1 (2008).

[269] SPIER, R. The history of the peer-review process. *Trends Biotechnol. 20* (2002), 357–358.

[270] STALLMAN, R. What is copyleft? Free software foundation, 1996. At http://www.gnu.org/copyleft/copyleft.html.

[271] STALLMAN, R. About the gnu project. Free software foundation, 1998. At http://www.gnu.org/gnu/thegnuproject.html.

[272] STALLMAN, R. M. *Free Software, Free Society: Selected Essays of Richard M. Stallman*. GNU Press, 2002.

[273] STRINGER, M. J., SALES-PARDO, M., AND AMARAL, L. A. N. Effectiveness of journal ranking schemes as a tool for locating information. *PLoS ONE 3*, 2 (2008), e1683.

[274] SUBER, P. Open access in 2007. At http://www.earlham.edu/ peters/fos/newsletter/01-02-08.htm.

[275] SUBER, P. Open access overview:focusing on open access to peer-reviewed research articles and their preprints [eb/ol] [2009-01-13]. At http://www.earlham.edu/.

[276] SUBER, P. Harvard law school joins harvard fas in mandating oa. *Open Access News - News from the open access movement* (2008). Available at: http://www.earlham.edu/ peters/fos/2008/05/harvard-law-school-joins-harvard-fas-in.html.

[277] SUNDÉN, J. *Material Virtualities*. Peter Lang, New York, 2003.

[278] SURE, Y., BLOEHDORN, S., HAASE, P., HARTMANN, J., AND OBERLE, D. The swrc ontology - semantic web for research communities. In *Conference on Artificial Intelligence (EPIA)* (2005).

[279] SWAN A., BROWN S. Authors and open access publishing. *Learned publishing 17*, 3 (2004), 219–224.

[280] SYMONDS, M. R., GEMMELL, N. J., BRAISHER, T. L., GORRINGE, K. L., AND ELGAR, M. A. Gender differences in publication output: Towards an unbiased metric of research performance. *PLoS ONE 1*, 1 (2006), e127.

[281] TAKACS, G., PILASZY, I., NEMETH, B., AND TIKK, D. On the gravity recommendation system. In *Proceedings of KDD Cup and Workshop* (2007), pp. 22–30.

[282] TAKACS, G., PILÁSZY, I., NÉMETH, B., AND TIKK, D. Investigation of various matrix factorization methods for large recommender systems. In *Proceedings of 2nd Netflix-KDD Cup and Workshop* (2008).

[283] TALLMO, K.-E. The misunderstood idea of copyright, 2005. At http://www.nisus.se/archive/050902e.html.

[284] Technorati blogosphere report 2007. Available at `http://www.sifry.com/alerts/archives/000493.html`.

[285] TENOPIR C.,KING D.W. *Towards electronic journals: realities for scientists, librarians, and publishers*. Special Libraries Association Washington, DC, 2000.

[286] THORNE F.C. The citation index: another case of spurious validity. *Journal of Clinical Psychology 33*, 4 (1977).

[287] TODD, P. A. AND LADLE, R. J. Hidden dangers of a 'citation culture'. *Ethics Sci. Environ. Polit. 8*, 1 (2008), 13–16.

[288] TOM GRUBER. A translation approach to portable ontology specifications. *Knowledge Acquisition 5* (1993), 199–199.

[289] TORVALDS, L. Tech talk: Linus torvalds on git, 2007.

[290] ULF ASKLUND AND LARS BENDIX. *The Unified Extensional Versioning Model, Lecture Notes in Computer Science Volume 1675*. Springer, 1999.

[291] UNDERWOOD, A. J. It would be better to create and maintain quality rather than worrying about its measurement. *Mar. Ecol. Prog. Ser. 270* (2004), 283–286.

[292] UREN, V., BUCKINGHAM SHUM, S., BACHLER, M., AND LI, G. Sensemaking tools for understanding research literatures: Design, implementation and user evaluation. *Int. J. Hum.-Comput. Stud. 64*, 5 (2006), 420–445.

[293] VAN DER AALST, W. M. P., TER HOFSTEDE, A. H. M., WESKE. *Business Process Management: A Survey*. Springer Berlin, 2003.

[294] VAN RAAN, A. F. J. Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics 62*, 1 (2005), 133–143.

[295] VAN RAAN, A. F. J. Comparison of the hirsch-index with standard bibliometric indicators and with peer judgement for 147 chemistry research groups. *Scientometrics 67*, 3 (2006), 491–502.

[296] VAN ROOYEN, S., BLACK, N., AND GODLEE, F. Development of the review quality instrument (rqi) for assessing peer reviews of manuscripts. *J. Clin. Epidemiol. 52*, 7 (1999), 625–629.

[297] VAN ROOYEN, S., GODLEE, F., EVANS, S. BLACK, N., AND SMITH, R. Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomised trial. *BMJ 318* (1999), 23–27.

[298] VAN ROOYEN, S., GODLEE, F., EVANS, S., SMITH, R., AND BLACK, N. Effect of blinding and unmasking on the quality of peer review: a randomized trial. *JAMA 280*, 3 (1998), 234–237.

[299] VELTRI, G. Social computing and the social sciences. Tech. Rep. TGE Adonis Report, Institut Jean Nicod, Paris, 2008.

[300] VINKLER, P. Characterization of the impact of sets of scientific papers: the Garfield (impact) factor. *JASIST 55*, 5 (2004), 431–435.

[301] VINKLER P. Evaluation of the publication activity of research teams by means of scientometric indicators. *Current Science 79*, 5 (2000), 602–612.

[302] VV.AA. *Assessing the impact of open access. Preliminary findings from Oxford Journals*. Oxford University Press, 2006. At http://www.oxfordjournals.org/news/oa_report.pdf.

[303] W3c organization. At http://www.w3.org/XML/.

[304] WALSH, E., ROONEY, M., APPLEBY, L., AND WILKINSON, G. Open peer review: a randomised controlled trial. *Brit. J. Psychiat. 176* (2000), 47–51.

[305] WARE, MARK. Scientific publishing in transition: An overview of current developments. At http://www.stm-assoc.org/storage/Scientific_Publishing_in_Transition_White_Paper.pdf.

[306] WARREN, J. C. Correspondence. *SIGPLAN Notices 11*, 7 (1976), 1–2.

[307] Web ontology language- wikipedia. At http://en.wikipedia.org/.

[308] WEINGART, P. Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics 62*, 1 (2005), 117–131.

[309] WELLCOME TRUST. Costs and business models in scientific research publishing, 2003. At http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtd003184.pdf.

[310] WELLMAN, B. *Structural analysis: From method and metaphor to theory and substance*. Cambridge University Press, Cambridge, UK, 1988, pp. 19–61.

[311] WENNERAS C.,WOLD A. Nepotism and sexism in peer-review. *Nature 387* (1997), 341–343.

[312] W.M.P. VAN DER AALST. The application of petri nets to workflow management. *The Journal of Circuits, Systems and Computers 8*, 1 (1998), 21–66.

[313] W.M.P. VAN DER AALST, MATHIAS WESKE, AND DOLF GRÜNBAUER. Case handling: A new paradigm for business process support. *Data and Knowledge Engineering 53*, 2 (2005), 129–162.

[314] Xml- wikipedia. At http://en.wikipedia.org/wiki/Xml.

[315] YU, B. AND SINGH, M. Searching social networks. In *Proceedings of Second International Joint Conference on Autonomous Agents and Multiagent Systems* (2003), pp. 65–72.

[316] YUZHONG QU, GONG CHENG, HONGHAN WU, WEIYI GE, XIANG ZHANG. *Seeking Knowledge with Falcons*. Springer, 2008.

[317] ZACHARIA, G. *Collaborative Reputation Mechanisms for Online Communities*. Massachusetts Institute of Technology (MIT), 1999. Masters thesis.

[318] ZHANG, Y.-C., BLATTNER, M., AND YU, Y.-K. Heat conduction process on community networks as a recommendation model. *Phys. Rev. Lett. 99*, 154301 (2007).

[319] ZHANG, Y.-C., MEDO, M., REN, J., ZHOU, T., LI, T., AND YANG, F. Recommendation model based on opinion diffusion. *Europhys. Lett. 80*, 68003 (2007).

[320] ZHOU, T., REN, J., MEDO, M., AND ZHANG, Y.-C. Bipartite network projection and personal recommendation. *Phys. Rev. E. 76*, 046115 (2007).

[321] ZIMAN J.M. Author public knowledge: an essay concerning the social dimension of science. *Cambridge: Cambridge University Press, 1968* (1968).

[322] ZIVKOVIC B. AND DERISI, S. Plos journals sandbox: A place to learn and play. *PLoS ONE 1*, 1 (2006).

[323] ZUCKERMAN, H. AND MERTON, R. K. Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system. *Minerva 9*, 1 (1971), 66–100.