SEWPS

SPRU Electronic Working Paper Series

## Paper No. 94

# Applying the Open Source Development Model to Knowledge Work

**Juan Mateos Garcia and W. Edward Steinmueller**
(SPRU)

**June 2003**

SPRU Science and Technology Policy Research

**The Freeman Centre, University of Sussex,**
**Falmer, Brighton BN1 9QE, UK**
**Tel: +44 (0) 1273 877155**
**E-mail: W.E.Steinmueller@sussex.ac.uk**
**http://www.sussex.ac.uk/spru/**

# Applying the Open Source Development Model to Knowledge Work

**Juan Mateos Garcia**

**W. Edward Steinmueller**

**SPRU – Science and Technology Policy Research**

**University of Sussex**

**INK**
**Open Source Research**
**Working Paper**
**No. 2**

**January**
**2003**

Information, Networks & Knowledge Research Centre

SPRU Science and Technology Policy Research

**Abstract**

This paper introduces a distinction between two different types of information goods in order to analyse the processes governing the review and integration of multi-authored contributions to information goods such as those produced through collaborations using the Internet as well as modular information goods such as open source software. It is argued that these distinctions are important because they suggest different organisational arrangements for producing such information goods.

This method of analysing the nature of the information goods is employed to examine different organisational arrangements using the analogy of collaboration for traditional publication to identify actors and processes. The analysis of 'contributors' is extended from authorship to collectors and researchers. The paper examines a small survey of the governance procedures employed in projects that employ open source methods for collecting various types of information. We noted the prime role of the recruitment process in the relative success of the examples that we examined (ODP, Wikipedia, Nupedia, MathLearning, VRoma, and Web of Life). For these 'collection' efforts, the role of hierarchy in editing and review of project submissions appears to be important than in open source communities and may be an impediment to recruitment and project development. A number of directions for further research are identified.

Keywords: open source software, collaboration, hierarchies, trust, teams, co-operation.

JEL Classifications: L86, L96, L33, P13, Z13

Acknowledgements:
This research project has been funded under National Science Foundation Grant NSF IIS-0112962, for the period 15 August 2001 – 31 July 2003. The able editorial assistance of Cynthia Little is also gratefully acknowledged.

# 1 Introduction

The extensive recent attention given to the Open Source Movement (OSM) can be seen as a re-appraisal of the opportunities for collective action in the creation of knowledge aided by the new infrastructures offered by the Internet. In recent years, there has been a steady increase in the frequency and amplitude of claims that markets based upon voluntary exchange of commodities for money are a universal solution to the problems of social organisation. This chorus has obscured the growing significance of other forms of organisation in modern industrialised economies including the non-profit sector and the part of this sector based upon voluntary initiative for the common good.[1] It also risks misleading conclusions being drawn about the commercial viability of market organisation of activities that, traditionally, were predominantly planned and financed by the public sector, such as education, scientific research, and public security. In this context, it is astonishing that one of the largest and most profitable industries of the 'new economy' may be disrupted by the appearance of communities of volunteers aiming to substitute public goods for proprietary commodities.[2]

Evaluation of the opportunities and limitations arising from the OSM is a most urgent task. On the one hand, the methods that this movement has established for collaborative endeavour suggest a new paradigm for the division of labour. At the extreme, this is a paradigm in which the modularisation of tasks is so profound that the spontaneous and voluntary contributions of individuals might create a common, collective, or public good that might rival of the efforts of hundreds or even thousands of employed software designers and engineers. On the other hand, there is substantial evidence that the coalescence of individuals in the communities or projects within this movement is an uncertain process, failing to ignite much more often than it captures and inflames the imagination and efforts of participants.

The broader relevance of the OSM is the possibility that it might be a model that is applicable to a much broader range of activities involving the creation of common or public information goods. The implementation of a global information infrastructure through the Internet and the World Wide Web offers unprecedented capacities for assembling large virtual communities, supporting collaborative endeavour, and distributing the results of these endeavours. These endeavours need not be limited to software. A much broader array of collective or public information goods are candidates for production using the methods pioneered by the OSM.

This paper examines the opportunities for extending the methods of organisation and collaboration utilised within the OSM to the production of a broader range of authoring and publishing, the creation and distribution of information. This examination is based upon a variety of evidence, conjecture, and reasoning that lead to an analytical or conceptual map that is partially validated using existing evidence from open source software communities and from other communities adopting 'open' methods for the creation of public or quasi-public information goods. It is argued that this conceptual map is useful for those planning to undertake or to fund virtual collaborations. Finally, the conceptual map introduced in this paper suggests several specific empirical investigations that may improve the value of this model in planning and implementing virtual collaborative communities

The paper is organised in four sections. Section 2 examines the structure of tasks and activities involved in the production of collective information goods. It identifies roles and considers how these roles are distributed in traditional and open source information-goods creation and publishing contexts as well as introducing the classes of collaborative collection and research to the analysis. Section 3 applies the framework discussed in Section 2 for the case of collaborations that are principally aimed at creating *collections* of information rather than for producing *information systems* in which the elements

---

1 The voluntary services sector is sometimes called the 'third sector' in Europe to designate its growing importance relative to the public and private 'for-profit' sectors.

2 Microsoft Corporation has joined with many market observers in noting the challenge offered by the OSM. The 2002 Annual Report observes 'The Company continues to face movement from PC-based applications to server-based applications or Web-based application hosting services, from proprietary software to open source software such as the Linux operating system, and from PCs to Internet-based devices.' (Microsoft Corporation 2003)

are tightly related to one another, e.g. software where one software module is required for another to work or scientific data where one observation may profoundly affect the search for other observations. The empirical examination of information systems (or what in this paper are termed vertical information goods with cumulative dependency) is not undertaken here for reasons of length.[3] Section 4 derives some conclusions from the examination, drawing out the implications for planning and funding virtual collaborative communities and indicating priority areas for future research.

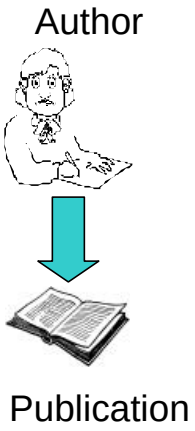
[3] A related approach, examining the specific issues of intellectual property governance in the context of collaborative research, is pursued in Cassier and Foray, 1999.

## 2 Organising the Collective Production of Information Goods

The simplest scheme for producing an information good as illustrated in Figure 1, is the single author publishing an information output. This simple diagram, however, abstracts from the processes that are 'black boxed' within the arrow - in the pictured case, the setting of the author's manuscript into type, its printing, and binding into a book.[4] We will ignore the publication infrastructure throughout this section as we wish to focus on the processes by which content is created and modified by collective processes. Thus, even though a single individual rarely performs all of the tasks of author-publication, it is quite common for all of the content to be prepared by a single individual.

The World Wide Web (WWW) has dramatically expanded the opportunities for self-publication, and a substantial fraction of the 3 billion web page that are currently indexed by the Google search engine have been produced by individual authors. While at first glance self-publication on the WWW can be seen as a solitary activity, it is significant that individuals often choose to include links to other individuals or to organisations that they expect to be of interest to their audience. Thus, even though an individual 'page' on the WWW may be self-published, its content may well embed by reference other information outputs, produced by other individuals or by the types of collective processes that are discussed in this section. This broader sense in which the WWW, itself, is a collective information good is also ignored in this section.[5] Our aim is to focus on the process by which individual web pages and their 'local' resources (resources maintained by the same author or group responsible for the web page that is linked to them) are created and maintained.

**Figure 1 The Single Author**




Our use of the term 'information good' is meant to include, from the universe of every possible arrangement of binary data, those arrangements that may be seen as having value to one or more individuals. In the simplest case, individuals may value the ability to record and recover their writing, recorded speech, etc. at some future date.[6] In the cases that are of primary interest in this paper, the aim is to exchange information either in the process of development or refinement of an information good that is meant for some group of final users. This group of final users may be confined to those participating in the process, it may extend to others (perhaps those willing to pay a fee for access), or it

4 In the tradition of William Blake, there are of course individuals that perform all of the tasks related to book publication. An even larger number of individuals purchase the services of a book publisher, submitting images of the book pages for mass reproduction and binding.

5 The production of search engines, guides, and directories to the World Wide Web is already a large business, indicating one of the ways in which new 'business models' may emerge from the free distribution of information, i.e. public information goods.

6 See Steinmueller 2000.

may comprise anyone that is interested in accessing the information good.[7] These cases, in economic terms, are instances of private, quasi-public, and public information goods. Examples of information goods whose exchange we will consider in this paper include texts and audio-visual material that is encoded in ways that can be displayed or 'played' by WWW browsers. These information goods have a wide variety of applications.

Although there is, in principle, no technical distinction between 'private' and 'public' information goods, i.e. they are both arrangements of binary information, the institutional rules governing the exchange of information as well as the possibility of encryption make such a distinction possible. These institutional rules, including copyright as applied to software or other rules such as the European Data Protection Directive, subject those who reproduce protected information goods to civil and/or criminal penalties. 'Private' information goods are therefore commodities that have 'owners' whose rights are, in principle, legally enforceable. It is also possible to make information goods private by encrypting them and limiting distribution of the means for their decryption.[8] Public information goods have no institutional rules preventing their exchange or reproduction by whoever gains access to them. A consequence of the capability to reproduce public information goods is that it may be difficult to receive a payment in exchange for their provision. Between these two forms of institutional rules, property rights and public information goods, a variety of other arrangements are possible, including some that permit the derivation of private goods from public goods and some that seek to ensure that public goods remain freely accessible. Our principal focus in this paper is on public information goods, but we will make a few observations about the implications of our analysis for the production of private information goods and other arrangements in passing.

With these preliminaries attended to, the remainder of this section focuses on how larger information goods production processes may be organised, beginning with a simple extension of the model introduced in Figure 1 aimed at accounting for many of the activities involved in traditional methods of publishing information goods. The model is developed in the next sub-section (2.1). This discussion is followed by a second sub-section (2.2) that explicitly considers how roles are assumed in open source software projects as well as how such projects govern offers of initiative and maintain coherence within the community. A third sub-section (2.3) extends the model one step further to consider two new classes of information good production, collaborative collection and research, which add additional actors and introduce further possible modifications to the model of information good publishing developed in the first two sub-sections.

---

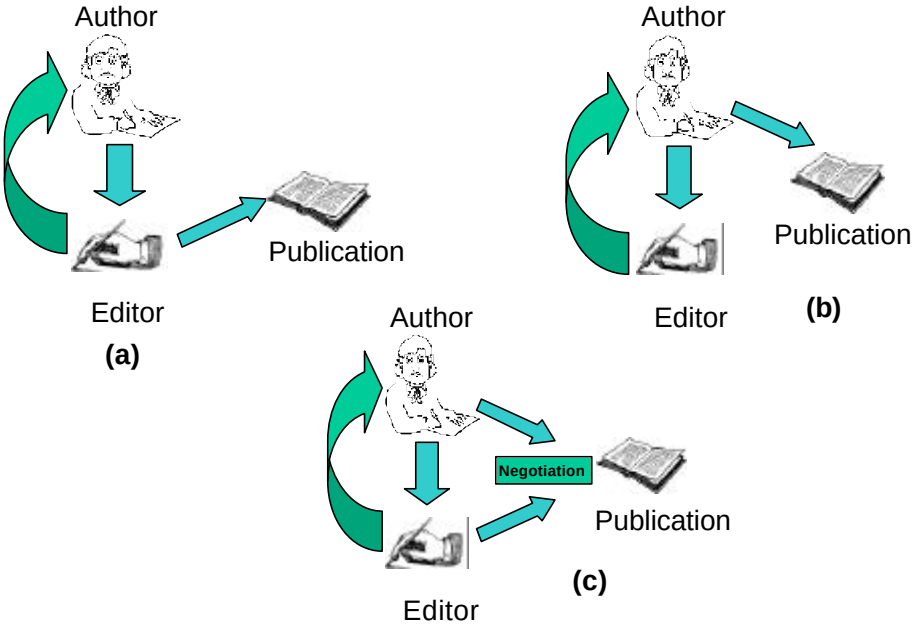
[7] It is common to speak of software products as commodities due to the nature of the legal protections provided to their producers which appears to make a given 'copy' of a software program a discrete unit of property. In the case of packaged software that is sold 'as is' and for which little or no support is provided, the idea of commodity is appropriate. If, however, it is packaged software with extensive user support or one of a large class of software that is meant to be customised and extended by the user such as the products produced by SAP or Oracle, then it is possible to make a choice in terminology. In one terminological scheme, the services that accompany the product may be recorded as after-sales support and the right to use the software is governed by a licensing arrangement. The alternative is to view the purchased software copies, the updates to which the user may be entitled, and all of the after-sales support services (including additional 'bits' to customise the user's installation of the software or add functionality to it) as a package. In this case, it may be more appropriate to view software as a 'club good,' a set of services and entitlements to which one may buy 'subscription' access. A considerable share of software that was previously produced as 'shrink wrap' packaged software is now distributed through authorised copying on networks, i.e. through a subscription arrangement. Much like audio-visual media, software is distributed within a system of 'digital rights management' in which users are bound by contractual agreements requiring them to maintain records of the copies made and in use.

[8] This can be done by prohibiting the reproduction of decryption information or, more commonly, by embedding the means for decryption in a technological artefact, such as a software system, which is controlled by the information good owner. Since decryption involves reconstituting the original unencrypted content of the information good additional measures must be employed to prevent the re-distribution of this content.

## 2.1 Roles in the Traditional Model of Information Goods Publication

Figure 2 illustrates three different configurations of authority or control in the production of an information good by the dyad of editor and publisher. The introduction of an editor into the information good production process raises issues of authority and control. In particular, who declares the manuscript to be finished and ready for publication? In part (a) of the figure, the author submits output to the editor who in turn may pass it directly to the publication process without further input from the author or may choose to return the edited work to the author for further revision prior to publication. In part (a) of the figure, however, it is the editor that determines when the work is complete. In part (b) of the figure, the author submits output to the editor and receives the revisions. The author then has the choice of sending it on for publication as edited, of revising it further (including over-ruling the editor's interventions) and sending it for publication, or sending it back to the editor for further suggested revisions. In this case, it is the author that has final control over the publication. In the final case, depicted in part (c) of the diagram, the author and editor agree to share the decision about when the information good is to be published.

**Figure 2 Configurations of Authority in Edited Publication**




Each of these assignments of publication authority have implications for the behaviour of the author-editor partnership. In configuration (a) where the editor has control there is the risk that the author will lose interest in the final product and the editor must win the author's trust that the interventions are in the author's best interest. Examples include cases where the editor has greater experience with the publication process than the author and is able to claim that the changes are necessary to conform to applicable standards or legal or market requirements. In the case of software, the editor may have a greater responsibility for maintaining the part of the system to which the author's contribution is directed. It is also possible to think of these configurations as principal-agent relationships. In the case of configuration (a), it is the author who is effectively the agent of the editor who must have particular incentives with respect to the publication process. In configuration (b), the editor is the agent of the author who remains the principal, responsible for the content of the final publication. This case will apply when it is necessary to retain the author's full commitment to the process. While the services added by the editor may be highly valued by the author, the author has discretion over whether or not specific interventions are included in the final manuscript. In the case of written publication, it is rare to give the author this much control, although this is sometimes done with literary works by established authors. Finally, in configuration (c), the author and editor agree to an explicit negotiation to determine

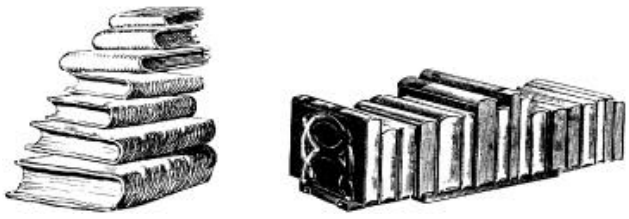
when the work is released for publication. This has the advantages of both of the prior configurations, but raises the problem of what happens when the author and editor cannot agree. In effect, both parties can veto the final production of the information good. Such possibilities suggest further rules concerning mediation or arbitration in the case of disputes.

This simple example illustrates how rapidly the problems of authority and conflict enter into the process of information good production. Once we depart from the very simple model of Figure 1, these problems of authority and conflict emerge and require the creation of social and/or legal institutions, i.e. rules, norms, and standards, to govern the process.

Applying this analysis to the case of collaborative production of public information goods further complicates matters. If both the author and editor have the intention of eventually publishing the information good, either may choose to release it for publication at any stage of the collaborative editing process. The ultimate users are faced with the problem of distinguishing which is the authentic or most appropriate of the information goods that have become available. Since such confusion is not likely to be in the interest of either the author or the editor, there is an incentive to resolve disagreements and achieve a negotiated 'closure' of the information good for publication.

Up until this point we have taken the information good being produced as a single output with little or no relation to other author-editor partnerships. Many information goods are, in fact, related to one another. Articles in an encyclopaedia, modules in a curriculum, or sections of a computer program are all examples of information goods that may be individually authored but that are meant to fit together into a larger assemblage or 'system'. Two fundamental types of such assemblages are identified in Figure 3. The vertical assemblage depicted on the left is meant to represent a relationship of 'cumulative dependency', the books (information goods) that are located at the bottom are pre-requisites for understanding the books that appear above them. As the stack grows, the extent of specialisation narrows and the need for additional information is reduced (the smaller size of the volumes appearing towards the top of the stack). By contrast the assemblage depicted on the right combines volumes of different sizes in a looser or complementary relationship. We will assume that the collection is not random, but that each of the items collected is meant to complement one or more of the others,[9] and will call this assemblage 'complementary dependency'.

**Figure 3 Basic Types of Information Assemblage**



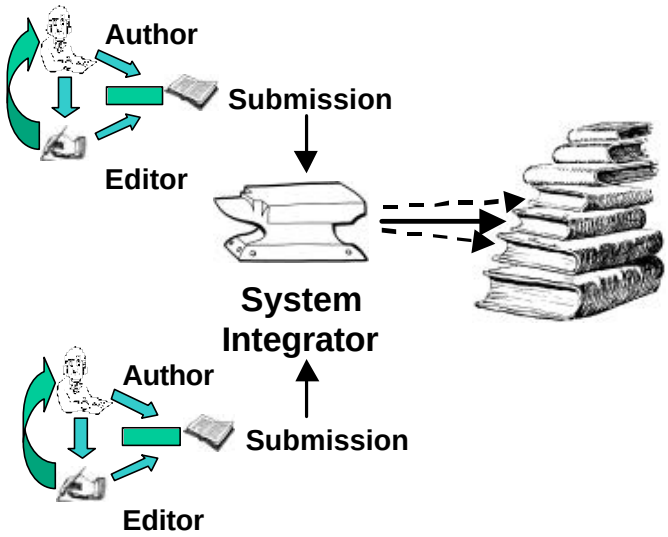

Different organisational arrangements are necessary to produce the components (individual information goods) of vertical and horizontal information assemblages because of the differences in the extent of

[9] Technically, we will want to have the entire collection connected in some way so that the separation of any single volume or sub-group of volumes into a different assemblage would reduce its total value to one or more users. In addition, we will want to hold that any scheme for separation does not lead to a set of sub-assemblages that have equal or greater value for all users. That is, we want to rule out the possibility that value might be generated through division into other assemblages to compensate for the loss of value in the primary assemblage. The aim here is to define the conditions for an 'integral' collection and it is assumed that 'complementary dependency' requires an integral collection.

inter-dependence between the components. In vertical assemblages, a component near the bottom of the stack that has missing or incorrect information will make the components near the top useless. By contrast, missing or incorrect information reduces the complementarity (and hence the value) of the horizontal assemblage, but it continues to be useable.

The role of the 'editor' is therefore likely to be substantially larger in the case of vertical assemblages than in horizontal assemblages of information components. Indeed, it is unlikely that authors would be able to make effective contributions to vertical assemblages if specific 'rules' regarding the interfaces between different levels were not specified before they begin their activities. Even with such rules, it may be necessary to make substantial revisions in the contribution before it can be effectively 'integrated' into the vertical assemblage. The function of performing this 'fitting' is not necessarily best done by the editor, whose talents may be better deployed in examining the internal consistency and completeness of the information good component that is meant to be integrated. Instead, a next step in the division of labour is to create a role for a system integrator. We employ an 'anvil' icon to represent this role, which of course may be conducted by either an individual person or a closely co-operating group of persons. The addition of the 'system integrator' is depicted in Figure 4. In this figure, the system integrator is depicted as needing to take account of the immediately adjacent information components in the assemblage in order to preserve the integrity of the entire system. This is only one possibility, as it may be the case that the 'dependency' spans more levels than simply the one that is most adjacent. The sole responsibility for integration of 'submissions' into the assemblage is depicted here as relying upon the system integrator. Other organisational schemes are possible that may give author/editor partners greater initiative to introduce their outputs more directly into the assemblage. Such arrangements may hasten the process of assemblage at the risk of losing coherence, a breakdown in the vertical dependence of the assemblage.

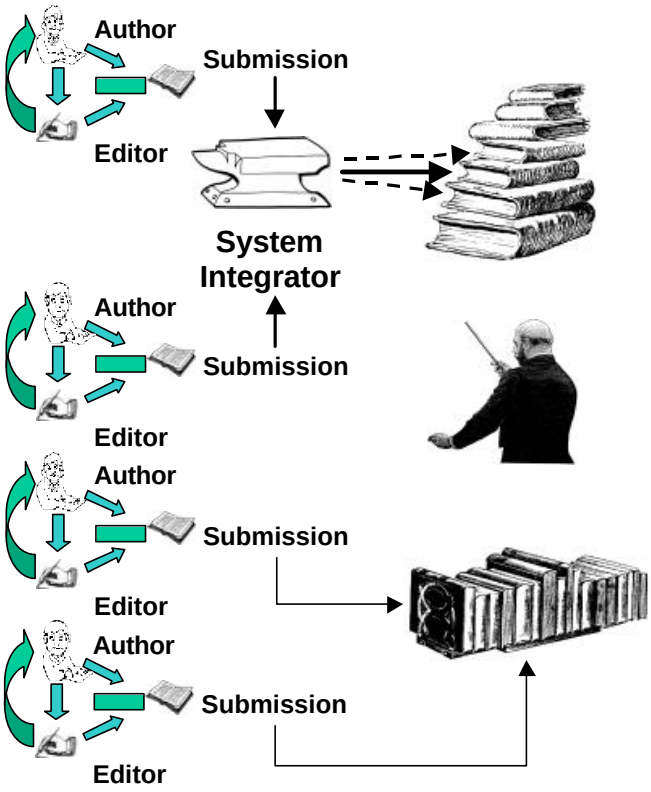
**Figure 4** **Adding the System Integrator in the Assemblage of Vertical Information Goods**




The system integrator function may also exist in the case of vertical information assemblages, but is likely to take a much greater variety of forms. For example, the system integrator may issue guideline standards for author/editor teams and only periodically or episodically check submissions for their adherence to these standards.

For either vertical or horizontal information assemblages, it is useful to include one more function, which is that of overall co-ordinator or 'conductor' who takes responsibility for the project as a whole meeting the needs of its ultimate users. The conductor is likely to be necessary because of the difficulties in getting the other actors to 'internalise' a common vision of the assemblage. At the same time, it would add substantially to the circuit nature of the assemblage process to have the conductor 'direct' or 'approve' the actions of author-editors or system integrators. The conductor is introduced in Figure 5 which depicts both the vertical and horizontal information assemblage processes as we have discussed them so far.

**Figure 5 Introducing the Conductor to the Information Assemblage Process**




While somewhat abstract, the discussion to this point has recapitulated roles that are recognisable from a wide variety of information good production processes involving multi-author collaboration. The roles and responsibilities of authors, editors, system integrators, and conductors may differ substantially between different categories of information goods assemblages as well as within categories of assemblages such as the compilation of curricular materials, encyclopaedias, edited collections, periodicals, some types of recorded music, software, and scientific databases. No distinctions peculiar to 'open source' methods for the division of labour have been introduced and no reference has been made to the distinct roles that individuals may play in such communities. The remainder of the section is devoted to examining how this traditional process may be altered by introducing features that are distinctive to the open source software movement and that can be reproduced in the context of other information good production processes. As we will see, the basic elements of the traditional method for division of labour do not disappear. Instead, some roles and processes are altered in minor ways, a new set of actors is added, and it is necessary to more carefully consider the mobility of individuals within these structures, i.e. their capacities to undertake different roles as they gain more experience with the project.

## *2.2 Division of Labour in the Open Source Method of Information Good Assemblage*

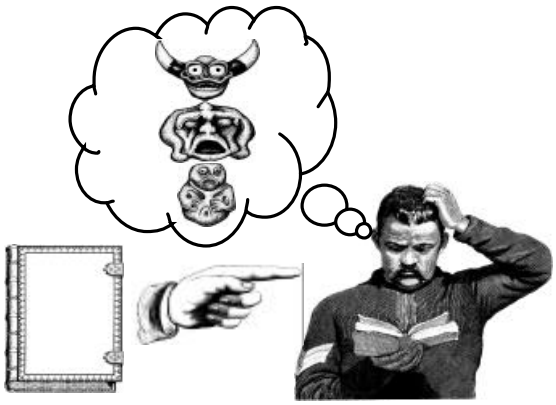
Consideration of the open source process begins with the issue depicted in Figure 6. The person receiving an information good assemblage from the process described in the previous section (2.1) may well be puzzled. This does not necessarily mean that the information good is 'bad' or 'wrong.' It may simply be that the individual is unfamiliar with the principles of its construction – they have no 'code book' to help them make use of the assemblage.[10] This is very likely to be true if they have not been involved in the writing or publishing of the assemblage. The user may need intermediaries (personal interaction or texts) to understand the potential use and value of an information good – e.g. they may want to read the 'user's manual' that is provided with software. Alternatively, after a period of head scratching as in Figure 6, the user may come to some understanding of the information good assemblage. In the case of computer software, this often occurs by installing it and experimenting with the user interface.

**Figure 6 The User Receives the Information Assemblage**


The more interesting possibility is that the user develops a critical understanding of the limits to the software. They find errors, omissions, or obscurities that reduce the value of the program to them and likely to other users as well. This perception is depicted in Figure 7. The user's response to these perceptions is important. One the one hand, they may simply reject the information good because of its perceived defects. Alternatively, they may choose to communicate to one of the individuals or groups responsible for the information good production. If subsequent 'editions' or 'releases' of the information good are produced, there is a possibility for the correction of these perceived shortcomings in the information product. Whether this happens depends not only the existence of subsequent 'versions' of the information good, but also on whether any of individuals engaged in the publication process are willing to listen to the user's suggestions about changes or modifications. Obviously, this will vary greatly depending on the objectives and incentives of the individuals involved and the clarity of the communication between the users and the producers of the information good.

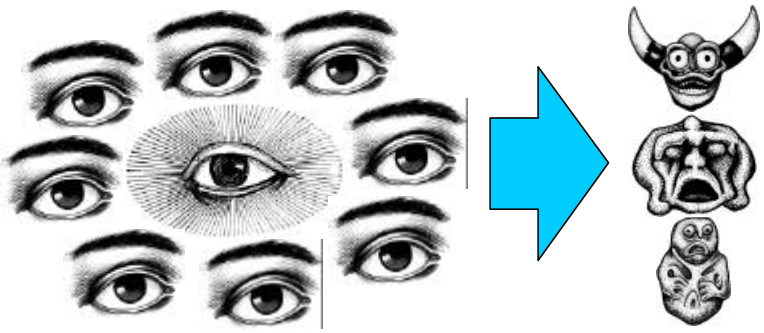
[10] The reference to 'code book' is drawn from the literature on knowledge codification which employs this term to identify recorded information that may help a user learn the meaning (i.e. gain a knowledge or understanding) of messages written employing standards, conventions, or norms meant to facilitate the reconstitution of knowledge from information. A simple example of a 'code book' is a dictionary of technical terminology which explains in simpler language the meaning of terms of art that might appear in technical journals or instruction manuals, see Cowan, David and Foray 2000.

**Figure 7 User Perceiving Errors, Omission and Obscurities**



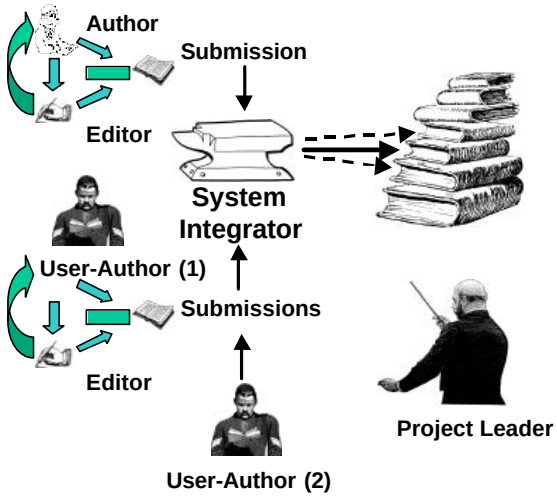
A major advantage that is claimed for the open source software development methods is their ability to have many different 'eyes' search for errors, omissions, and obscurities in software, the information good produced by this community. What distinguishes open source communities from other 'reader suggestion' or 'user suggestion' types of arrangements is their active encouragement of individuals who identify deficiencies to offer a 'fix' or 'patch' intended to correct the perceived problem. In software production, this is both a significant hurdle and an opportunity. By recruiting some users as producers, open source communities refresh and extend their membership, increasing the vitality of the effort and the possibility for innovations to be introduced from outside the original community of developers (von Hippel 2002). At the same time, however, the simple acceptance of user alterations, such as software patches or fixes, bypasses the editor and system integrator functions and introduces the risk that the information product will lose coherency.

**Figure 8 The Eyes Have It – Identifying Errors, Omissions and Obscurities**



Thus, while the user may well become an author in the system, it is also necessary to preserve integrity of editorial and system integrator functions. We have hypothesised in an earlier paper (Mateos-Garcia and Steinmueller 2003) that the means of achieving this is through 'distributed authority,' a negotiated access to control of the design, integration, and editing of open source work in progress. In open source software communities, the creation of distributed authority, a hierarchy in the design, integration and editorial function of a project is of particular importance because of the 'vertical' nature of the product being produced and the relative wide span of inter-dependency between different elements of the information product. This is the feature that we identified earlier in this paper as cumulative dependency. The result of embedding user-authors in the system described earlier in Figure 5, is depicted in Figure 9. Two possibilities are illustrated. In one, the user-author (1) works with an 'editor' to produce a submission to the system integrator and in the other the user-author (2) makes a submission directly to the system integrator. It is of course also possible that either the editor or the system integrator may be user-authors, although this possibility is not depicted in Figure 9.

**Figure 9 Embedding the User-Author in the Information Assemblage Process**




The idea of distributed authority can be understood as the ability of entrants, such as the user-author, to move to other roles in the project. Even the project leader's position is not sacrosanct, as this person or group may lose interest in the project and transfer control to other individuals including individuals that entered the project by becoming user-authors.

In summary, the advantages of the open source way of working arise from the principles of 'the eyes have it' –many individuals inspecting a project's source code and making proposals for fixes or patches in response to their perceptions of errors, omissions and obscurities. In making such suggestions, these users are encouraged to become user-author(s) as depicted in Figure 9 and by entering the system gain influence over the project. The distribution of authority evolves and changes with the growth of the project and in response to the perceived value of the author and user-author's contributions by the 'more senior' members of the community. Indeed, the distinction between author and user-author is an artificial one reflecting how the individual was recruited to join the community. The author is an individual that joined the project with the prospect of future use while the user-author is an individual that joined after attempting to use a version of the software, or information assemblage in the more abstract framework. Since open source communities do not seem to get very far without a working version of the programme, virtually all authors are likely to have some features of the user-author.

It is worth preserving this distinction however, as a way of keeping track of the distinction between the self-interest and collective interest motives of the participants. We may expect that the individuals we are calling authors may have a more flexible view of the implementation of the project, being primarily committed to aims other than their specific needs including, perhaps, the widespread use of the information good. The user-author, on the other hand, is motivated in the first instance by a desire to introduce specific changes relevant to his/her needs and is likely to have less of an interest in how widespread is the use of the information good. The flexibility that we have suggested is a distinguishing characteristic of the open source way of working makes it possible for user-authors to assume a wider vision of the collective use of the information good. It also makes it possible for authors to become user-authors, interested in shaping the evolution of the product towards their own specific needs.

This discussion raises several implications for the processes of governance of these communities. A significant problem that the community may face is the introduction of major revisions by either authors or user-authors. Such revisions may not accord with the self-interest or the perceptions of what is the collective interest of a large number of individuals in the community. Such revisions raise the possibility that 'hijacking' or 'forking' of the project will occur (Raymond 1999; Mateos-Garcia and Steinmueller 2003). In a 'hijacking,' the individuals favouring the revision depose the project leader who has resisted the revision, leaving this original leader with no followers. In a 'fork', two distinct implementations emerge, one that remains with the original project leader who does not accept the

revision, and the other with a new project leader (who is likely to be the individual or group introducing the major revision). One means of governing the possibility of forking is to follow the heuristic, release early, release often as this practice makes implementation of a major revision difficult. In the time needed to make a major revision, a series of incremental changes will have occurred, creating a 'moving target' for the sponsors of the revision.

A defensive strategy that can be adopted by the project leader is to pre-emptively delegate authority to the author or group sponsoring the revision. This will be more likely if the revision sponsor's aims do not fundamentally conflict with the aims of other parts of the community including the project leader's 'vision' of the aims and goals of the project. Neither of these strategies will work against an individual or group that is sufficiently determined to alter the course of the project and can win a substantial 'following' within the community. Although it is not yet clear from the evidence about the operation of such communities, it is possible that this prospect may influence the size or complexity of viable open source development efforts, making smaller and more modest efforts vulnerable to revisions and hijacking or forking behaviour.

This section has examined the publication practices of the open source community and, at the same time, the parallels that these practices have with other forms of publishing activities such as book, scientific journal, newspaper, or encyclopaedia publishing. The principal distinction to be drawn between these different types of activities is whether they in the main involve a vertical and cumulatively dependent assemblage of information (e.g. modules in a software system) or a horizontal and complementary dependence on other information goods (e.g. articles in an encyclopaedia). In situations where cumulative dependence is dominant, it is expected that role of 'system integrator' will become essential. In traditional organisations, each of these roles can be assigned to specific individuals and the resulting team resembles the 'surgical unit' described in Brooks (1982). Two features distinguish an open source software community. A first feature is the flexible and fluid allocation of roles. This feature is the result, in part, of the practice of recruiting those who are capable of making useful improvements in the software project (including the remedying of errors, omissions, and obscurities) and, in part, of the implicit challenge to established roles that this recruitment brings. Authority is 'distributed' as a means of preventing the hijacking or forking of the project. A second feature of the open source community is the openness that results from the publication of source code, which encourages the entry and recruitment process of the first characteristic as well as the growth of a community of users who are knowledgeable about the product. The first feature is difficult for proprietary software developers to imitate because of the tensions that would be introduced by admitting developers from 'outside' into the active life of a software development process. For example, who would be responsible for the time lost in considering submissions and how could an unpaid volunteer be allowed to serve as a system integrator? The second feature may be partially imitated by proprietary software developers if arrangements can be made to open source code to key or 'lead' users and to prevent the re-distribution of the source code to those wishing to avoid purchasing the software. Similar issues face the organisation of any knowledge work using open source principles.

At this point, we have only considered two processes of information creation, authorship and editing (taking system integration to be a type of editing). The next section considers other processes of information processing that are, in the first instance, most relevant for the creation of horizontal or complementary information goods (assemblages) – the roles of the collector acting in collaboration with others and of the researcher. These new roles extend the possible sources of information that might be organised according to open source principles and illustrate the elasticity of the open source model to the organisation of other knowledge work tasks.

## 2.3 Two Additional Types of Information Good Creation: Collaborative Collection and Research

Information does not come only from the author's imagination; it also comes from the external world. The variety of life on this planet, the stars and other objects in the sky, and the immense accumulation of human creations (artefacts, texts, images) are examples of features of the external world that have information content and that have motivated the efforts of individual information 'collectors' over the generations. While many collectors amass information as a private good during their lifetimes, it is eventually released as a consequence of their mortality and often becomes a public inheritance for later generations. Other collectors amass information on behalf of human kind (or social groups) sponsored by the state or philanthropic organisations. Curators of museums and librarians are amongst the large tribe of individuals that serve as information collectors for society. The implications of collection as a method of creating 'horizontal' or complementarily dependent information submissions to an assemblage are considered in the first part of this sub-section. The latter part of the sub-section is devoted to examining the issues involved in collaborative research, a process of information generation that shares some of the 'vertical' or cumulative dependency features of software creation as well as some of the 'horizontal' or complementary dependence features of collection activities.

**Figure 10 The Collaborative Collector**

Figure 10 is an iconographic representation of a cluster of collaborative collectors engaged in a common 'search' for information that will become 'submissions' of the type discussed in the previous sub-sections. The distinguishing feature of *collaborative* collectors is that they will operate with some degree of co-ordination, i.e. they will in some way take account of others' actions. This co-ordination may be directed at avoiding the duplication of effort or at dividing the labour of the collection process. It may also involve the 'alignment' of their efforts to increase the extent of complementary dependency in the creation of a horizontal information assemblage. In this case, much of the organisational apparatus discussed in the previous section will be relevant.

The development of the World Wide Web has led to the creation of the over 3 billion URLs (which may be considered 'pages' of information) and 8.7 million sites.[11] To put these numbers into

[11] Based upon Google's search engine, which is only capable of accessing publicly accessible URL addresses, see http://www.google.com/ and http://www.google.com/press/facts.html (last accessed 15 February 2003) for the current statistic and the definition of 'pages' as URLs. The size of the World Wide Web as measured by the number of sites is from the Online Computer Library Center, Inc. (OCLC), http://www.oclc.org (last accessed 15 February 2003). This total includes sites that are not publicly accessible or are 'provisional,' defined by OCLC as being 'under construction' or containing content that would be 'trivial' to the general user. OCLS reports that the number of publicly accessible sites is 3.08 million as of 2002, see 'Web Characterization Project' at http://www.oclc.org/research/ (last accessed 16 February 2003). Comparing the number of pages given in the text to this number of publicly accessible sites indicates that there are some very large sites (as measured by pages), that Google's count of the number of URLs may include URLs that would not ordinarily be accessed as sites (e.g. ftp directories), or both.

perspective, the US Library of Congress, the largest library in the world, contains some 18 million books and 118 million items in total. The World Wide Web is about the size of the Library of Congress' book collection if we imagine that the average size of a book volume in the Library of Congress is somewhere near 170 pages (i.e. 18 million * 170 = 3,060 million or 3.06 billion). Other collections exceed the size of the Library of Congress in terms of total items, but have smaller book collections.[12] In terms of the number of individual items, however, some of the world's largest databases contain larger horizontal information collections. For example the largest astronomical database includes 1 billion individual objects,[13] while the human genome project aims to obtain the DNA sequence of the 3 billion DNA sub-units present in human DNA.

Human interest, however, is not limited to the amassing of very large collections that require public support. Examples of smaller collections that are documented on the WWW include barbed wire,[14] fad objects from the past (e.g. Rubik's Cube),[15] and Roman coins.[16] Many such sites are 'sole author' or semi-commercial efforts – primarily motivated by an interest in sharing with others, perhaps with some small financial return to help defray the costs of providing the information assemblage on the WWW.[17]

The WWW clearly encourages the sharing of information about collections. It is not clear how successfully collection activities can be organised using the principles of open source collaboration. The main issue is developing a common perspective on the reason for the compilation of the collection which will motivate individual contributors to adhere to some common standard that makes donating their efforts to a collaborative effort more desirable than self-publishing. In other words, the contributor to a collaborative collection will have to value the complementary dependency of their contribution relative to others. In the next section (Section 3), we will provide some examples of collection efforts that have begun on the Web. At the time of writing, however, only a very few efforts appear to have achieved the sort of critical mass that is likely to sustain them indefinitely.

A second additional type of information creating activity that may be suitable for open source type methods is 'research,' represented iconographically in Figure 11. In the first instance, research involves the collection of findings, traditionally the domain of academic publishing, the compilation of scientific databases, and the interpersonal exchange of research data and communications. This facet of research publication closely resembles collaborative collection activities. The principal distinction is that there is likely to be a greater collective interest in quality control of submissions. One reason for this is the issue of 'priority,' the credit for first discovery of new scientific knowledge.[18] Another is that a principal use of such collections of research results is as input into other scientific research. Faulty inputs will result in waste of time and resources in scientific research or in efforts to apply scientific knowledge to technological problems. Both of these issues suggest that the role of 'editor' and 'system integrator' will assume a larger function in efforts to organise collaborative research resources.

---

12 For example, the British Library counts individual philatelic items, see http://www.bl.uk/about/annual/pdf/perfindic.pdf (last accessed 15 February 2003). Both the British Library and the Bibliotheque Nationale de France contain in the vicinity of 10 million books.

13 http://www.projectpluto.com/datasets.htm#A2 (last accessed 15 February 2003).

14 http://www.barbwiremuseum.com/ (last accessed 15 February 2003).

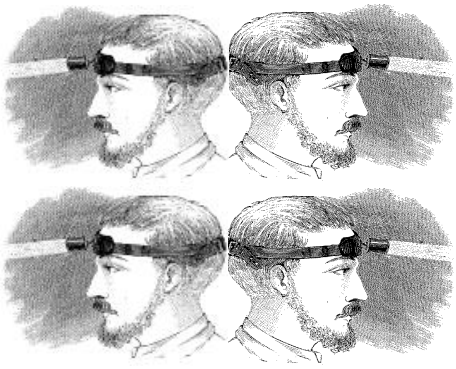
15 http://www.badfads.com/home.html (last accessed 15 February 2003).

16 http://myron.sjsu.edu/ (last accessed 15 February 2003).

17 The rental of server space is likely to be the dominant cost for many such efforts.

18 Merton (1957), Dasgupta and David (1994)..

**Figure 11 The Researcher as Information Good Assemblage Author**



In addition, the potential for using scientific findings to produce other scientific findings suggest that the contributors to such information collections are likely to delay their publication in the hopes that new ideas for exploiting the data will occur to them. With time, however, the recall of particulars necessary to understand and make use of the data will fade, making the contribution useless. Moreover, there is a bottleneck in the process of contributing scientific data and results in that researchers are likely to be paid to generate findings and disseminate them in traditional forms, but not to prepare and enter them in a collection. Some research funding authorities require deposition of data as a condition of funding, but there is unevenness in the funding for the costs of properly archiving data and very little professional reward for improving the quality of such submissions.[19]

These issues suggest that there will be significant problems in compiling scientific contributions in horizontal collections. As we will see in the next section, creating information assemblages comprising research results is not an activity that has, so far, achieved a high level of participation. This indicates a very important point that in our view offers considerable validation for the model suggested here. It is not that, as a general observation, scientific research databases are few in number or shallow in depth. These problems only arise with *horizontal* collections of scientific research results. The databases that are actively compiled in the *process* of scientific research are *vertical*, i.e. their outputs are expected to be inputs into further scientific research. These vertical information collections include the data from the human genome and other genetic sequencing efforts or streams of data from high-energy physics and atmospheric and oceanic studies. Because of the range of externalities offered by these data collections, researchers have an incentive to develop the social institutions necessary for validating data (the editing and system integration functions) before it is embedded or for distinguishing between preliminary and 'confirmed' data, features that we would expect to see in information assemblage activities with a high degree of cumulative dependency. This analytical conclusion has several important implications for the planning and implementation of future programmes for information assemblage that will be explored in the conclusion to this paper (Section 4)

This sub-section has introduced collaborative collection and research as two additional activities that might lead to submissions to information assemblages. Collaborative collections may be expected in a large number of areas of human interest. The principal problem they appear to face is deriving a vision for a 'horizontal' structure that will be valued by users. Complementary dependency requires a collector to conform to a submission standard, whilst in the context of the global reach of the Internet, a collector may be as, or more, attracted by the opportunities the WWW offers for self-publication. Examples of collaborative horizontal assemblages that have enlisted significant participation are provided in the next section along with a partial analysis of the processes that support their collaborative character. Higher hurdles appear to be involved in horizontal compilations of scientific results due to concerns about quality control of such information when the 'entry' process is open. We would argue that the efforts required in validating and organising scientific data are most likely to be made when such databases are part of a *vertical* information assemblage process.

[19] The case of economics is considered in David and Steinmueller (1991).

A principal conclusion of this section is that vertical information good production provides the strongest motives for collaboration because the possibilities of sharing in the externalities generated by other participants are matched by incentives to edit and integrate contributions due to the cumulative dependency of contributions. Horizontal efforts, involving complementary dependence, are neither so well developed nor so likely to lure participants from the opportunities offered by self-publication and looser affiliation (e.g. links to other websites). It is not clear at this stage of development whether these shortcomings are the result of failure to develop compelling visions for structuring horizontal information assemblages that would raise the level of value derived from complementary dependency and ignite more widespread and intense participations. Alternatively, it may simply be that open source methods in software have been particularly intense because of the immediate value of these vertical information assemblages and the facility of the principal contributors in the use of the Internet as a communication and collaboration method. The next section provides some insights into these competing explanations as well as offering examples of open source methods employed for creating information goods that are not software code.

# 3 Open Source Methods without Code: Applying the Principles to Horizontal Knowledge Collection

This section investigates examples of WWW sites for evidence about the ideas introduced in the preceding sections. The claim made at the conclusion of the last section was that we would expect to see substantial differences in the intensity and depth of activities involving vertical as opposed to horizontal information assemblage. In addition, it was suggested that horizontal information assemblage activities would benefit if there were methods for strengthening the extent of complementary dependence of contributions. In this section, we begin with some of the most successful examples of horizontal information assemblage collaborations and proceed to discuss the less successful. This leads us to the final section (Section 4), which offers some conclusions and implications for future research.

As noted earlier, it is possible to see the WWW as a whole as a kind of collaboration to create a horizontal information assemblage. In terms of the framework discussed previously, the WWW involves entirely optional participation of editors and system integrators. Additions are made to the assemblage with or without such interventions and it is not therefore surprising that a very large share of all publicly accessible pages are 'under construction' or offer content viewed as trivial by the general reader. The Online Computer Library Center, Inc., which compiles statistics on all IP addresses that respond to a request for WWW page delivery, reports that for 2002, the share of such addresses that provide free, 'unrestricted access to all or at least a significant portion of its content' was 35%, while the share of sites that 'in transitory or unfinished state (e.g., "under construction"), and/or offers content that is, from a general perspective, meaningless or trivial was 36%.' The balance, 29%, was comprised of 'private' sites that offered little or no useful (accessible) information.[20]

In other words, given a valid address for a Web page server, a user has only about a one third chance of finding information that could conceivably be of use. Sorting amongst the 3.08 million public sites for information that would be useful for a specific user is a major challenge.[21] For example, an individual spending 10 seconds on each site would take over 20 years to survey the entire content of the Internet.[22] Fortunately, this problem is addressed by a number of different 'search engines' that utilise various schemes for indexing the WWW's content and providing a means for users to search using key words and other search conventions.[23] At least one of these search engines, Yahoo!, uses a mixture of automatic methods and human editors. Yahoo! and other search engine providers offer a number of ways for site maintainers to provide information that will make their sites appear near the top of the search results or appear in response to a wider array of key words. In this sense, the search engine providers are engaging in a collaboration with site maintainers to produce a horizontal information assemblage and the providers engage in different types of editorial efforts in this sort of collaboration. This kind of collaboration is organised according to commercial principles in that the search engine database remains under the sole control of the search engine provider.

The most prominent example of collaborative information assemblage involving open source principles is the Open Directory Project (ODP), which claims to have enlisted the aid of over 55 thousand editors and to have classified some 3.8 million WWW sites into 460 thousand categories.[24] In the framework that we have developed here, ODP operates according to open source principles as a community of contributor-editors enrolling to edit the listing of WWW sites in specific subjects according to their

---

20 'Private' sites are defined as '[sites whose] content is intended for a restricted audience; restriction can be explicit (e.g., fee payment or authorization) or implicit (obvious from nature of content).' See 'Web Characterization Project' at http://www.oclc.org/research/, last accessed 16 February 2003).

21 See Steinmueller (1992) for an analysis of sorting and searching issues.

22 Assuming a working year of 250 days and an 8 hour working day leads to the answer 23.4 years.

23 See http://www.google.com/about.html (last accessed 16 February 2003) for links to a discussion of how one of the leading search engines, Google, achieves this using automated methods.

24 See http://dmoz.org/ (last accessed 16 February 2003).

interests and expertise. The ODP conforms quite strictly to the technical principles of the open source software movement. Its entire content is downloadable and usable by others, the only restriction being that explicit recognition must be given to the ODP. All of the software necessary to operate this database is available in source code and can be freely downloaded as well. The social organisation of the ODP is explicitly hierarchical. A person enlisting in the project must apply before submitting content and the application is included by a group of 'senior editors' whose identities are unpublicised but can be found relatively easily. One of the project's co-founders remains active in the project and serves as project leader or conductor and the project is 'owned' by Netscape, Inc.

The ODP has a complete structure for the distribution of authority with seven categories of editors with increasing levels of responsibility. Recruitment begins with completion of an application. The guidance for applicants about the review of their application is summarised as follows:

> Your application will be reviewed by one or more senior ODP editors, who will evaluate it on the basis of the answers you provide. Since no two applications are the same, there is no magic formula for an acceptable application. Each application is treated uniquely. We make every effort to accept people who will add value to the directory and the community, however, we don't accept all applications. You may reapply if your application is denied.[25]

The ODP can be seen as a 'fork,' i.e. a division in the historical development path of a project or line of research, in the effort to create human edited websites in response to the problems that developed at Yahoo! in processing site applications as summarised by the following comment on the early history of the ODP:

> But as Web growth explodes, critics have charged that up to a third of the sites seeking a listing in the popular Yahoo directory don't make it in. The company has admitted that some sites may take months, even years, to get listed at all.[26]

The ODP was established with the name GnuHoo, a deliberate reference to the Gnu (Gnu's Not Unix) free software community, changing its name to NewHoo and then to DMOZ or the Open Directory Project. It is a clear example of a successful effort to employ open source ways of working in the context of a collaborative information assemblage and is notable not only for its size, but also for the fact that it is owned by a corporate sponsor.

Arguably, the feature that makes the ODP successful is that it achieved a sufficiently large scale to attract contributors. Shortly after its launch on 5 June 1998 it claimed to have 400 editors covering 3,900 categories. The categories were, in turn derived from a hand edit of the usenet (a bulletin board system that was developed before personal computers and remains active today) interest groups that had evolved as a reflection of user interest. From 400 editors to 55,000 editors in four and a half years is a measure of the success of this and other decisions made in structuring the community according to open source principles. While there are seemingly endless collections of link resources available on the Internet, i.e. self-published versions of the Web guides, the ODP offers the competitive attraction for volunteers of being able to tap the substantial externalities available from the single information assemblage of the ODP. Despite its relative success, however, the ODP does not have a particularly strong attraction for search engine users and does not appear amongst the top ten search engines in various recommendations of which search engines to use. In this sense, the ODP has yet to achieve its ambition of displacing some of the leading search engines.

A dramatically less hierarchical collaborative collection project is the Wikipedia, an open source encyclopaedia project founded by Jimmy Wales, also the founder of Nupedia (discussed briefly below). Wikipedia employs all of the principles of open source including the Free Software Foundation's Gnu licence, which requires that works incorporating open source content shall also remain open.

Wikipedia allows users to edit virtually all of the pages and provides for a relatively simple transition from being a contributor to having more privileges to alter content and organisation of the horizontal information assemblage that it represents. Since its founding 15 January 2001, Wikipedia has

---

25 See http://dmoz.org/cgi-bin/apply.cgi?where=Society/Philosophy/Philosophy_of_Science.

26 The Search Engine Report (1998), http://www.searchenginewatch.com/sereport/98/07-newhoo.html, (last accessed 16 February 2003).

succeeded in attracting 8,300 registered users who have produced over 100,000 articles, a slightly larger number than the Encyclopaedia Britannica makes a claim to.[27] A distinguishing feature and difficulty with the Wikipedia is that because anyone can instantly become an editor, some topics are controversial and some articles may not be entirely reliable.

What Wikipedia has gained from opening the editor function to all users is a community with strong and self-generated norms including a working understanding of a key term 'neural point of view' that is frequently mentioned on the site. All users are free to make changes, but this includes the prospect of changes being rejected in favour of earlier versions. Thus, individuals that make changes that are not *valued* are relatively quickly discouraged. More serious vandalism is dealt with similarly with the ultimate sanction of blocking the originators IP address, a sanction that has been relatively rarely applied.[28] In practice, the etiquette of editing involves outlining changes that are to be made and seeking comment before proceeding to make major changes. Surprisingly, this process seems to work rather well. The main shortcoming appears to be that there are many topics that have not attracted the level of activity necessary to transform them into relatively mature texts.

Comparing the success of Wikipedia with the uncertain future of Nupedia is instructive. [29] Nupedia, also a collaborative encyclopaedia project, is founded on many of the same principles as Wikipedia but is based upon a well-articulated peer review and editing model similar to the one described earlier in this paper. Nupedia may have been responding, in part, to the criticisms about the absence of a validation or authentication process on the Wikipedia. This is explained in the introductory page to Nupedia in the following way:

> This project seems so rigorously edited that I feel I might not be welcome. Are you concerned about that?
> We most certainly are. We want everyone to feel welcome. Nupedia is a genuine online community, and we want you part of it. If you would prefer a more open project, you might, however, be more interested in Wikipedia, which we hope will complement Nupedia. Nupedia editors and peer reviewers are not associated (except by their individual choice) with Wikipedia; it's a separate project.[30]

The other distinguishing feature of Nupedia is the intention of eventually generating revenue through 'banner advertising' during the use of the site. The use of this revenue model and the perception (whether true or not) that author's contributions will benefit the Nupedia organisers may be a barrier to Nupedia's growth and the Nupedia site reports that the project is temporarily suspended while efforts to find funding for its further development are pursued.[31]

Two further examples of collaborative collection sites that have had a degree of success are noted to indicate the range of topics that may be suitable for organisation by open source methods. The first is ExploreMath, a site where members can contribute interactive mathematics lesson plans.[32] While there are already a large number of sites offering 'self published' (often by organisations) resources for teachers, the ExploreMath site is an attempt to generate the externalities possible through indexing and using a common set of tools for documents and visualisations. ExploreMath is a corporate site. Its parent company explorelearning is venture capital-backed and is engaged in developing a number of related activities, e.g. an ExploreScience site.[33] It remains to be seen whether some of the issues that have been barriers to Nupedia's growth will influence the future of the explorelearning venture.

A second example, Vroma, is a site that offers a virtual re-creation of ancient Rome where teachers (and others) can, in principle, contribute teaching materials and make improvements. The project was

---

27 http://www.wikipedia.org/wiki/Special:Statistics (last accessed 16 February 2003).

28 Based upon examining the list of blocked IP addresses reported at http://www.wikipedia.org/wiki/Special:Ipblocklist, (last accessed 16 February 2003).

29 http://www.nupedia.com/ (last accessed 16 February 2003).

30 http://www.nupedia.com/about.shtml

31 See http://www.nupedia.com/, which reports on the last accessed date 16 February 2003, that 'The project is largely "on hold" but poised to get moving again with renewed vigor when funding is found.'

32 http://www.exploremath.com/ (last accessed 16 Februay 2003).

33 See http://www.explorelearning.com/, (last accessed 16 February 2003).

initially funded by a $190,000 grant from the National Endowment for the Humanities and its futher appears somewhat uncertain. Signs of life include the upgrading of the software to an open source MOO (for Multi-User Dimension (or Dungeon) Object Oriented environment), a common tool for implementing a complex and interactive environment in which users can navigate a common virtual space (containing various scenery and objects) and interact with one another. On the other hand, it is clear that much of the original development community has disappeared in the post-grant phase of the project and it is not clear that the two new 'owners' of the project are planning to develop the 'community' efforts further. In this sense, the project may evolve in the direction of a self-published resource for teachers of Latin as the site provides a large number of links to specialised self-published sites on this topic.

Finally, it is worth considering an effort that offers a very grand vision, but seems to be having considerable problems in achieving critical mass despite some excellent initial efforts. The Tree of Life project organised at the University of Arizona (Tucson) is an ambitious attempt to document the entire phylogenetic structure of life on earth through the collaboration of biologists from all over the world. Contributors are organised according to a peer review system and clear arrangements have been made for preserving intellectual property in matters such as photographs of specimens and copyright of texts. The project appears to have received about three quarters of a million dollars in funding from various sources including the National Science Foundation.[34] It has benefited from the efforts of more than 320 biologists from 21 countries who have provided very high quality contributions.[35]

In general, however, not very much of the tree of life has been populated. For example, in the case of modern birds (neornithes) the category of perching birds (passeriformes) is well-documented with a wonderful article on the class characterised as 'the largest and most diverse commonly recognized clade of birds.'[36] However, the well-organised tree diagram leading downward towards the individual species is not populated with any of this very large number (circa 10,000) of individual species (like crows or robins). Nor does the higher level classification of clades in which passeriformes appears have any other clades described (thus omitting discussions of penguins, woodpeckers, pigeons, etc.).

Thus, this project, which is the oldest of all of those described in this sub-section has not progressed very far in terms of the vision that it established.[37] No doubt this is due in no small measure to the limitation of contributors being professional biologists who may well receive better recognition and compensation for contributing their knowledge to scientific journals and field guides. In this respect, it is interesting to speculate what would happen if a more open policy concerning contributions were to be established and all of the amateur biologists throughout the world were invited to join in the effort, focussing the efforts of the professionals on editing submissions or identifying submissions as 'pending review.'

This section has provided a brief survey of collaborations aimed at creating horizontal information assemblages. The features of the collaborations examined are generally consistent with the model offered in Section 2. It appears important for the success of a project that the specific role of the contributors, editors, and system integrators be well-defined and that the collaboration take an active role in recruiting participants. It is also apparent that attaining a threshold level of engagement or involvement is critical to the success of such efforts. This is so in the first instance because the principal value of collaborative horizontal information assemblages arises from the externalities of having a considerable amount of related information at one location, integrated with a common format. It is clear from the experience of ODP that a strong hierarchy or corporate ownership is not, in itself, a fundamental barrier to the success of such efforts. Instead, what appears to be true is that the recruitment mechanism is the most critical element. Wikipedia is very successful in recruitment despite the uncertainty that accompanies all submissions (the strong editorial role given to other members of the community) and the absence of a hierarchy to govern individual submissions. Nupedia and the Tree of Life projects both appear to suffer from a relatively limited recruitment of contributors despite their intention to produce and achievements in producing high quality content. These same communities indicated that a strong vision and a high quality 'seed' (initial implementation) do not appear to be

---

34 http://tolweb.org/tree/home.pages/funding.html (last accessed 16 February 2003).

35 http://tolweb.org/tree/home.pages/history.html (last accessed 16 February 2003).

36 http://tolweb.org/tree?group=Passeriformes&contgroup=Neornithes (last accessed 16 February 2003) by Mr. Scott V. Edwards of the Burke Museum, University of Washington.

37 This project had its origins in work done in the 1990s.

overridingly influential on the success of subsequent efforts.[38] Finally, this section briefly considered a two further examples (ExploreMath and VRoma) of collaborative activities that are rather young[39] and are still attempting to sort out the issues of recruitment and control. They are offered as suggestions that somewhat more specialised activities than link resources and encyclopaedias may be suitable for open source type organisation. These examples offer a glimpse into a series of stimulating research issues and a first set of conclusions about the relation between the nature of the information good and the organisation for its production that are examined in the next section.

[38] (Raymond 1999) argues that these are central features in the case of the vertical effort represented by open source software.

[39] The first is young because of its recent birth, while the second may be considered young because the current project leaders inherited the project after a period of National Endowment for Humanities funding.

# 4 Conclusions and Implications for Future Research

This paper has made a distinction between horizontal and vertical assemblages of information in order to analyse the processes of review and integration of candidate submissions to such assemblages. The term assemblages has been employed to capture the multi-authored character of information goods produced through collaborations using the Internet as well as the modular nature of these information goods. Horizontal information assemblages, also referred to here as 'collections,' were defined in terms of the by the nature of the complementarity dependency between the items.[40] Such collections are like a library or encyclopaedia. Their value is increased by the inclusion of more volumes or articles, but they would still have value if one volume or article were missing. Vertical information assemblages, also referred to here as 'systems,' were defined in terms of cumulative dependency among items.[41] Their value may be catastrophically reduced if a single item is defective (having errors, omissions, or obscurities). These epistemological distinctions are important because they suggest different organisational arrangements. In horizontal information assemblages, there is less need to 'police,' i.e. govern, the fit between components than in vertical information assemblages.

This method of analysing the nature of the information good was then employed to examine different organisational arrangements. In conjecturing about the governance requirements of the two types of assemblages it was useful to consider the nature of the collaborative process in authoring, editing and publishing and further analogies were drawn to the cases of 'authorship' that involved system integration (vertical assemblages), collector-participants (producing horizontal assemblages) and researcher participants (producing vertical assemblages). In our empirical work, we examined the governance procedures in projects involving collection or horizontal information assemblage. We noted the prime role of the recruitment process in the relative success of the examples that we examined (ODP, Wikipedia, Nupedia, MathLearning, VRoma, and Web of Life).

This finding concurs with one of the key findings of the literature on the open source community which also suggests that later stages of a project's life require 'scaling up' in the number of participants. However, a considerable variety of arrangements for distributing authority (a second feature of open source software communities) were discovered among the six examples examined. Both strong hierarchical authority arrangements and highly decentralised provisions of authority are identifiable from existing communities. This finding suggests that collaborative projects involving horizontal information assemblages may not need, and indeed might be harmed by, the articulation of authority arrangements that are necessary in collaborative vertical development groups. This finding suggests that further research among a larger set of collaborative communities involved in producing horizontal information assemblages would be warranted and useful. This paper also suggests that further examination of the establishment and evolution of means of distributing authority in vertical information good creation projects would be useful, a task that would have been undertaken in this paper if space constraints had not been an issue. In particular, are different arrangements suitable at different stages in the life of the project? The emphasis on recruitment also indicates the need for further examination of the 'life cycle' of both horizontal and vertical collaboration efforts for evidence on the relation between community size, hierarchy development, and project effectiveness.

Our continuing interest in the origins and evolution of authority as a means of governance indicates an urgent need to examine more closely the motives, work habits, and sponsorship of individuals that are active in information good creation. Such an examination should emphasise the mobility (or lack thereof) of individuals after they join in collaboration, and attempt to identify what factors are involved in individuals becoming more deeply engaged and active in such efforts. We have not considered in any detail the competition between projects for recruits or for the attention of information users nor

---

[40] In the terminology of economics, the items are 'weak' complements.

[41] In the terminology of economics, the items are 'strong' complements in the sense of fixed combinations of inputs. However, in this case, the economics terminology creates the potential for misunderstanding. There may, in fact, be a large number of equally satisfactory complements to an existing vertical assemblage that are substitutes for one another and the 'pieces' or modular components of a system may be assembled in a variety of ways. In this sense the strength of the complementarity is contingent on the nature of the implementation and some pieces may not fit with others.

have we examined the 'linkage' between different projects according to principles of operation, overlap of contributor or user communities, or use of particular tools to facilitate collaboration. These are clearly areas where further research is needed.

While considerable work remains to be done, this paper has suggested two guidelines that should be followed by those funding or intending to lead collaborative. First, when a project's success depends upon a large user community it appears to be useful to outline a sufficiently broad plan to accommodate the variety of possible interests of participants and to sketch in enough elements of this plan to convince contributors that there is some urgency in their making a contribution. Second, while an extensive hierarchy in the review of submissions provides a relatively certain assurance of the quality of the results it comes at a cost of discouraging entrants of authors and editors that may be critical to the project's success. In this sense, the aphorism 'recruit early, recruit often' might be added to those already in use in the open source community.

Finally, this paper suggests that open source principles are applicable to a wide variety of activities and that models for their application and adaptation are evolving as the experience in different projects is accumulated. This emerging experience provides a fertile ground for future research and a plethora of opportunities for new initiatives.

## References


Brooks, F. (1982). *The Mythical Man-Month*. London, Addison-Wesley Publishing Company.

Cowan, R., P. A. David and D. Foray (2000). "The Explicit Economics of Knowledge Codification and Tacitness," *Industrial and Corporate Change* **9** (2): 211-254.

Dasgupta, P. and P. A. David (1994). "Toward a New Economics of Science," *Research Policy* **97** (387): 487-521.

David, P. A. and W. E. Steinmueller (1991). "The Impact of Information Technology upon Economic Science," *Prometheus* **9** (1 (June)): 35-61.

M.Cassier and D. Foray (1999). "Les modes de régulation de la propriété intellectuelle dans les consortia de haute technologie," *Economie et Société, Série Economie de l'Innovation* **LII** (2).

Mateos-Garcia, J. and W. E. Steinmueller (2003) "The Open Source Way of Working: A New Paradigm for the Division of Labour in Software Development?" Falmer, UK, SPRU -- Science and Technology Policy Research, INK Open Source Working Paper No. 1. January.

Merton, R. K. (1957). "Priorities in Scientific Discovery: A Chapter in the Sociology of Science," *American Sociological Review* **22** (6): 635-59.

Microsoft Corporation (2003). *Annual Report 2002*. Redmond, Washington, Form 10-K, Item 1.

Raymond, E. S. (1999). *The Cathedral and the Bazaar: Musings on Linux and Open Souce by an Accidental Revolutionary*. Sebastopol, CA, O'Reilly and Associates, Inc.

Steinmueller, W. E. (1992). "The Economics of Production and Distribution of User-Specific Information via Digital Networks" in C. Antonelli, *The Economics of Information Networks*, Amsterdam, North Holland**:** 173-194.

Steinmueller, W. E. (2000). "Will New Information and Communication Technologies Improve the "Codification" of Knowledge?," *Industrial and Corporate Change* **9** (2): 361-376.

von Hippel, E. (2002) "Horizontal innovation networks - by and for users" Cambridge, MA, MIT MIT Sloan School of Management, Working Paper No. 4366-02. June.