TRUST AND EPISTEMIC COMMUNITIES IN BIODIVERSITY DATA SHARING

Nancy A. Van House
School of Information Management
and Systems
University of California
Berkeley, CA 94710-4600 USA
+1 510 642 0855
vanhouse@sims.berkeley.edu

ABSTRACT

Trust is a key element of knowledge work: what we know depends largely on others. This paper discusses the concepts of communities of practice and epistemic cultures, and their implication for design of digital libraries that support data sharing, with particular reference to practices of trust and credibility. It uses an empirical study of a biodiversity digital library of data from a variety of sources to illustrate implications digital library design and operation. It concludes that diversity and uncomfortable boundary areas typify, not only digital library user groups, but the design and operation of digital libraries.

Categories and Subject Descriptors

H3.7 [**Information Storage and Retrieval**]: Digital libraries -- *User issues.*

General Terms

Management, Design, Human Factors.

Keywords

Biodiversity; trust; credibility; communities of practice; epistemic cultures; users.

1. INTRODUCTION

> Knowledge is a collective good. We rely on others. ....The relations in which we have and hold knowledge has a moral character, and the word I use to indicate that more relation is *trust*... I argue that the identification of trustworthy agents is necessary to the constitution of any body of knowledge…[W]hat we know of comets, icebergs, and neutrinos irreducibly contains what we know about those people who speak for and about those things, just as what we know about the virtues of people is informed by their speech about things that exist in the world ([39]; italics in original)**.**

In all knowledge work, we rely on others – present and distant, known and unknown. Digital libraries have the capacity to radically alter practices of collaboration and knowledge work by making it possible to use, not only published work, but unpublished work from a variety of sources. Libraries, including digital libraries, that contain published information participate in the institutional structure of the publishing system that vets and warrants knowledge. Networked information, however, can cross the social and technical boundaries between published and unpublished, private and public, and local and global that used to, practically if not ideologically, limit our possibilities for sharing information knowledge work.

This increased access highlights questions of trust: Whom or what do we believe? How do we decide? And how do we design digital libraries to facilitate these judgments?

The social nature of information and of knowledge work and its implications for the design of computer-based systems are of concern in a number of research areas, including computer-supported collaborative work (CSCW), recommendation and collaborative filtering systems [31], and social navigation [13]. A growing literature investigates the assessment of Web-based resources (e.g., [1, 4, 7, 8, 35]. A new book [2] presents an array of socially-informed investigations of digital library use.

Most of these investigations are rooted in empirical observations of people's behavior or in system design projects, sometimes with reference to the literature of social psychology and organizational behavior. This paper, in contrast, draws on social epistemology ([10, 14, 52]) and science studies [17], two areas with long histories of investigating the social nature of knowledge. My contention is that their insights could fruitfully inform systems design. The dialogue could potentially go both ways. Digital libraries can make visible previously taken-for-granted practices of knowledge work, questioning and clarifying the understandings of social epistemology.

In this paper, I draw on the literature of social epistemology, situated learning, and science studies to better understand the practices of trust in knowledge work, the role of community in assessments of credibility, and design decisions in a digital library of biodiversity data. The purpose is not to present an ideal solution to problems of assessing credibility of sources, but to examine the practices by which people address this problem on a day-to-day basis by looking at how a specific digital library (DL)

JCDL'02, July 13-17, 2002, Portland, Oregon, USA.

addresses this problem – and to try to understand these practices in the light of epistemic cultures and communities of practice.

In another paper [48], I discuss the implications of these design choices for social epistemology. In this paper, I discuss the implications for the design and management of digital libraries. The major conclusion is that digital libraries operate both within and across epistemic communities. These communities differ in sometimes-subtle ways that must be considered in the process and content of digital library development. These differences affect, not only digital library users and content providers, but the people responsible for the digital library's creation and management.

2. BIODIVERSITY DATA AND NETWORKING

Sharing data has always been central to the work of science. Recent developments in Information and Communications Technologies (ICT) have increased the capabilities for data sharing, and raised new questions.

This paper is concerned with biodiversity research, which studies the diversity of life and the ecosystems that maintain it. Biodiversity research has complex data needs [3]:

> [Biodiversity research requires] communication and coordination — among agencies, divergent interests, and groups of people from different regions, from different backgrounds, and with different points of view. Biodiversity and ecosystem data can be politically and commercially sensitive and entail conflicts of interest. The kinds of data scientists have collected about organisms and their relationships vary greatly in precision and accuracy, and the methods used to collect and store these data are almost as diverse as the natural world they document. Many important observations are made by non-scientists, such as amateur birders and natural history enthusiasts. And the range of datasets with which these datasets must interact is unusually broad, including geographical, meteorological, geological, chemical, physical, and genomic sources. There is thus an unusual need to accommodate differences in data quality within a democratized community information infrastructure that is both formal and informal. ([30], p. 3)

An important source of fine-grained biodiversity data is observers in the field, many of whom are expert amateurs, with considerable expertise but no formal credentials. To use their work requires systems of recording, collecting, and collating their observations; evaluating their expertise; and assessing the accuracy of specific reports. Mistaken reports can sometimes have major repercussions. For example, is a report of a rare plant in an unusual location a significant sighting, or an identification error?

This reliance on amateur observers is not new. Fields such as botany, ornithology, and astronomy have long relied on such people. What is different now is the ease with which observations can be collected and reported and their volume. For example, the Audubon Society's Christmas Bird Count has used teams of amateurs for over a century. But now the Cornell Lab of Ornithology's Great Backyard Bird Count (http://birds.cornell.edu) is collecting bird observations over the Web on an on-going basis.

In earlier research on digital libraries, my colleagues and I [36, 46, 47, 49] investigated the concerns of environmental data users and producers about networked information. We found considerable enthusiasm about the potential of networking, but we also heard worries. Data users worried about assessing the quality of data available from unfamiliar sources, and problems with poor quality data. As the Internet makes it easier for people to publish on their own, by-passing systems of screening and evaluation, it places a greater burden on users to evaluate information.

Data owners reported increasing demands to make their "raw" data usable, not only to their colleagues, but more widely via the Internet. They expressed both enthusiasm and concern [47]. Environmental planning and biodiversity work are highly political, with major environmental and economic consequences. Some respondents were concerned that their work could be used inappropriately, due to either of two causes: technical incompetence; or inappropriate goals and values, such as when data were used to harm environmentally-sensitive areas.

3. CALFLORA: A BOTANICAL DIGITAL LIBRARY

This paper is concerned with issues of trust and credibility facing CalFlora (http://calflora.org), a nonprofit organization supporting a comprehensive web-accessible database of plant distribution information for California. Its goal is to provide ready access to data needed to identify critical issues in conservation of plant diversity, and to analyze consequences of land use alternatives and environmental change on distribution of native and exotic species. It serves researchers and the general public.

CalFlora is an independent non-profit governed by a board, with an Advisory Committee representing various contributor and user groups. CalFlora is currently hosted by the UC Berkeley Digital Library Project. Plans are underway for CalFlora to become free-standing when the project ends.

This paper is based on interviews with participants, observation at meetings, and review of documents.

For this analysis, we are interested in two components of CalFlora, the CalPhotos California plants and habitats photographs, and the CalFlora Occurrence Database.

The photo database, a joint project of CalFlora and the UC Berkeley Digital Library Project, contains over 20,000 images of California plants[1] from the California Academy of Sciences and other organizations, with a substantial portion from native plant enthusiasts not trained as botanists. Photo providers retain their rights, and each photo is labeled with a statement about allowable re-use. Photographers supply metadata. The typical record contains the scientific name of the taxon, date and location of the photo, institutional source (if any), and photographer's name and contact information, and links to occurrence records for the same taxon.

[1] The CalPhotos project contains other photos as well; the concern in this paper is the collection of plant and habitat photos accessible via CalFlora.

The occurrence database contains over 800,000 reports of observations of plants in California – specimens in collections and reports from the field – primarily from 19 institutions or organizations.

Occurrence records include taxon name, observer and contributing institution, date, observation type, documentation, location, and observer name. They are used to document the distribution of plant taxa around the state and changes over time. They are useful for increasing floristic knowledge, determining the current state of California flora, and for assessing the effects and potential effects of actions. An important part of land use planning is an assessment of the potential effect on the plant population, which requires comprehensive information about the area in question and comparative information for the rest of California.

Contributors of occurrence records and photos must allow free use of them by CalFlora users. However, records and photos will be removed from CalFlora any time the owners wish. (A set of state government photos were recently withdrawn from CalPhotos as a result of increased national security concerns.)

The major quality issue with both photos and plant occurrence records is taxon identification. Differences in plant identifications occur for several reasons. In some cases, the science has changed. Nomenclature changes over time; what was the appropriate identification when a record was added may be no longer. Outright errors are also a problem, particularly with the photos, many of which come from people without botanical credentials. Distinctions between species sometimes hinge on subtle characteristics which may not be apparent even to an experienced observer (e.g., differences visible only at the specific points in the plant's life cycle). The CalFlora website warns that, for photos, generally the genus is correct but the species may not be.

An annotation system for photos, described below, provides an avenue for users to submit corrections to taxon identifications. Occurrence records remain the property and the responsibility of the data providers, and users are asked to send updates or corrections directly to the institutional source. An annotation function is under development.

3.1 Working with Representations

The major strength of CalFlora is its large number of images and records and the linkages among them. Many observations collected across space and time are "synoptically present" [22] in standardized inscriptions and with associated tools that facilitate combination, juxtaposition, and comparison.

Much of scientific work consists of creating and using representations or inscriptions [23, 29]. Recent socially-informed discussions of representation in science emphasizes that inscriptions are not simply representations of the natural order, nor simple accounts of scientific work, but are themselves socially organized and contextually created and understood.

The observer functions as a scientific instrument, collecting, interpreting, and reporting. Anything that the observer fails to see, misclassifies, or fails to report appropriately is lost. Observers are trained in perception, coding, and inscription. Goodwin [15], for example, demonstrates how archaeologists learn the embodied practices of coding dirt. They learn to determine color by obtaining a specimen, squirting it with water, and comparing it to a standardized chart. They learn to distinguish how sandy a dirt specimen is by tasting it. Becoming an expert is learning to perceive, to categorize, and to inscribe.

CalFlora's representations vary in their distance from the empirical. Some records represent herbarium specimens that the observer could in theory inspect, although in practice most users do not. Photos provide visual evidence to support the observer's report, but are limited by such factors as magnification and perspective. An image of a field of yellow flowers near the sea, surrounded by sand dunes, may provide a good sense of the context but not enough detail to identify the species; a close-up of an individual specimen won't display much context.

With both photos and occurrence records, the user is dependent on the observer for information about the observation and the taxon identification. CalFlora users have to decide when to trust a representation; and CalFlora's designers have to provide users with the information that they need to make these judgments, to decide what and whom to trust and under what circumstances.

4. TRUST

Trust is a topic in, among other areas, philosophy [19], sociology [27, 44], and political science [11]. Many approaches to trust distinguish between cognitive and emotional factors [19]. Cognitive approaches divide into two kinds: those that consider trust as risk assessment, judgments of a person's or institution's competence and reliability; and those the focus on reliance, on one's disposition to act based on trust in another. Emotional factors focus on trust as a feeling rather than a cognitive assessment.

Another way to distinguish among notions of trust is to look at types of interpersonal exchanges or cooperation. One approach, typified by the papers in Gambetta [12], addresses division of labor, contracts, and exchanges. A second approach is concerned with sociability: the role of trust in the social order, civic engagement, and the relationship of trust to citizenship, cooperation, reciprocity, and morality (e.g., [9, 34, 38, 44]). Putnam's [34] and Fukuyama's [9] discussions of social capital and engagement in civic activity fall into this category.

The collective nature of knowledge foregrounds the type of trust of most interest in this paper, which has been called epistemological trust [5] or the granting of epistemic or cognitive authority [14, 52]. Wilson [52] distinguishes between expertise and cognitive authority. Experts are knowledgeable, but we grant cognitive authority to those whom we would ask for advice. His example is astrology: we might grant that a person is an expert astrologer, but not follow her advice.

Accepting others' testimony is, among other things, a strategy of cognitive efficiency. We have neither the ability nor the resources to make all possible observations, develop our own methods, and test all possible knowledge claims. Trust reduces transaction costs; in this case, the costs of developing or verifying knowledge claims on our own. Nor do we necessarily wish to: Wilson [52] points only that only a few knowledge claims are of sufficient importance for us to engage in detailed examination. He argues that we generally don't actually evaluate many claims; we wait until we need to decide whom or what to believe, and then weigh the costs of evaluating claims against the penalties of believing wrongly.

4.1 Trust and Information and Communication Technology

There is a small but growing literature on trust in computer-mediated communication, collaborative technology, and the Internet. Some discussions focus on trust in technology; others on trust in individuals.

One area of investigation is the design of trustworthy systems, systems that do what people expect despite environmental disruption, human user and operator errors, and hostile attacks [37]. A second is contract-like online agreements, that is, e-commerce and e-services [40]. This literature has been concerned with how people can engage in exchanges with people they may not know, and how systems can be designed to facilitate this cooperation (e.g., eBay).

A third area of research is concerned with the Internet and sociability. Some discussions consider trust in virtual communities and online social interactions, including presentation of self and identity and the possibilities for deception, hurt feelings, and embarrassment [6]. Another line of research asks about the effect of the Internet on "real" communities. Uslaner [45], for example, citing Putnam's concerns that people are increasingly disconnected, asks what role the Internet may play. Using data from the Pew Internet surveys he concludes that going online neither builds nor destroys trust, and that trusting people are no more or less likely to go online than misanthropes.

Another area of investigation addresses the effect of communication media on collaborative work, and designing technology to support collaboration within and across work groups and between individuals, such as advisor/advisee and agent relationships [18, 16, 32, 33].

Of particular relevance to this paper is credibility of Web resources, which addresses evaluative criteria for web sources, and the extent to which people use questionable information from the Web [1, 7, 8, 28, 35, 35] – that is, epistemological trust and networked information. Some of this work is descriptive; other is normative.

4.2 Bases for Judgments of Trustworthiness

What conditions promote trust? How do people assess one another's trustworthiness? Here we are mostly interested in epistemic trust, specifically how people decide to believe information from others and trust their work.

The discussions of the bases for trust are both empirical and philosophical. Competence and honesty are commonly named as primary criteria for trustworthiness [14, 52]. Competence is relative: we recognize degrees and spheres of competence [52].

We assess trustworthiness in a variety of ways. We rely on our prior experience with the person in question. Failing that, we rely on the experience or judgments of others whom we trust. For example, eBay compiles ratings of people's behavior by others. When we lack direct evidence, we rely on indicators of capability such as educational credentials or professional experience. We also rely on feelings of trust, caring, and familiarity. Informal social interaction and exchange of personal information promote feelings of trust [33].

Another set of proposed bases for trust are shared values, common cognitions, and social similarity [21]. We expect that people with whom we share values and understandings, people who are members of our group or like ourselves, will be trustworthy. And we look to our community for information about whom to trust and how to decide. A major function of professional communities is an on-going circulation of judgments about other members of that community [52].

The literature on assessment of web-based resources indicates that people's evaluation of information on the Internet relies heavily on their evaluation of the source [1, 7, 8, 28, 35]. Lynch [28] reports a new form of deception in the networked world: deceiving search engines to increase the ranking of a document or web site among retrieved results. He asks how we manage metadata in such environment. His answer is that, currently, the most reasonable solution is to determine the identity of the person or organization responsible for the metadata. He sees two possible ways to do this: a centralized, formalized approach to deciding what's including and what is not, which places great power in the hands of system designers -- and could easily devolve into censorship; or providing users with what they need to establish a source' s identity and determine their own willingness to believe information from that source. He terms this provenance.

Burbules [4], in a review of issues and methods of determining the credibility of online materials, concludes that "the Web is both an information archive and a social network; as people move within the space, their interaction with ideas and information is, at the same time, an interaction with other individuals and groups" (p. 450). He describes the networked environment as comprising "communities of obligation and commitment." He concludes that "in the end, the best safeguard is to check one's judgments against the judgments of the community with which one has confidence; choosing that reference group prudently is as much a moral matter, involving issues of respect and trust, as a matter of expertise" (p. 453).

Obligation, commitment, and similarity of values and cognitions are often attributed to culture, including national and organizational culture. However, such discussions generally black box culture, failing to ask what the word means or how culture functions in these assessments. In developing the notion of epistemic cultures, Knorr Cetina has opened the black box on culture and knowledge.

4.3 Epistemic Cultures

Science studies is deeply concerned with knowledge [17], and with how scientists come to decisions about what they agree to be true. Much of current of science studies subscribes to some version of the principle of symmetry: that for sociological analysis, one proceeds the same way in explaining beliefs that come to be seen as true and those that do not. In other words, "that it's true" is not sufficient reason to explain how groups come to decide what's true.

Knorr Cetina [20], a researcher in science studies, argues that, for all the discussion about contemporary Western society as a knowledge society, little attention has been paid to the nature of knowledge processes and the workings of expert systems. She introduces the concept of epistemic cultures, which she defines as "those amalgams of arrangements and mechanisms – bonded through affinity, necessity, and historical coincidence – which, in a given field, make up *how we know what we know*. Epistemic

cultures are cultures that create and warrant knowledge, and the premier knowledge institution throughout the world is, still, science" ([20], p. 1, emphasis in original). She claims that epistemic cultures are structural features of knowledge societies, and not limited to science.

The word "culture," she says, implies history and on-going events, attention to symbols and meaning, and, most of all, diversity: she argues against the naïve assumption of the unity of science. (She also explains why epistemic cultures are not the same as disciplines: the concept of discipline doesn't reflect "the strategies and policies of knowing that are not codified in textbooks but do inform expert practice" (p. 3).)

She roots her definition of culture in practice: the acts of making knowledge, and the dynamic patterns of activity. She is interested, not in the production of knowledge, but in the construction of "the machineries of knowing composed of practices," technical (e.g., scientific instruments) and social (e.g., how decisions are made).

She argues that these machineries are constitutive of knowledge and of the scientists and other actors. That is, epistemic subjects (variously the individual scientists and the collectives of them, labs and experiments) are shaped by, determined by, the practices and machineries of knowing. Her focus is on neither the knowledge produced nor the producers, but on practices that are constitutive of epistemic subjects and objects alike.

Knorr Cetina analyzes two epistemic cultures within the natural sciences, molecular biology and high energy physics. She uses the differences that she uncovers between them in their organization, practices, and understandings, their technical and social machineries, to demonstrate that even within science epistemic machineries differ.

So what's useful about the notion of epistemic cultures for digital libraries? First, it firmly situates knowledge in the social. Second, it emphasizes the amalgam of practices and mechanisms by which people do their work. It investigates epistemic machineries, the inner workings of expert systems. And most importantly, it emphasizes diversity and discontinuity: Knorr Cetina's argument is that epistemic cultures differ, even within science.

4.4 Communities of Practice

Knorr Cetina's emphasis is on the varieties of epistemic machineries. Another approach to knowledge, one that also emphasizes expert practice but is more concerned with the epistemic subjects, is Lave and Wenger's concept of communities of practice.

Lave and Wenger [24, 25, 26, 51] ask how new members are brought into knowledge communities, and how knowledge communities both transform and reproduce themselves. Theirs is a theory of situated learning that focuses on the person-in-the-world, as a member of a sociocultural community. Learning is not just receiving a body of factual knowledge; it is activity in and with the world; it is creation of identity. Person, activity, cognition, meaning, knowing, and world are interdependent.

Learning, then, takes place in community, and consists of becoming a skilled member of a community of practice. They deliberately avoid explicitly defining community of practice, but they do say that members of a community of practice share activity and understanding of the meaning of what they are doing in their lives and the world. They also say that "a community of practice is an intrinsic condition for existence of knowledge, not least because it provides interpretive support for making sense of its heritage" (p. 98).

What do communities of practice contribute to this investigation? Like epistemic cultures, the notion highlights situated activity; activity is not simply set in a context, it is mutually constituted with the context. Members of communities of practice share understandings about what they are doing and what it means, not just skills but orientations, values, and interpretations. The emphasis is on identity and mode of being. Finally, by contextualizing knowledge, the notion of communities of practice legitimates different knowledges [50] – like epistemic cultures, it warns us that different knowledge communities will have, not just different methods, but different epistemic machineries and understandings.

So we conclude that people from different communities of practice, different epistemic cultures, have, not only different methods of doing their work and determining whose work is "good," but diverse knowledges, understandings, ways of seeing the world and their role in it. Culture implies diversity.

We have also seen how members of communities of practice are constituted differently by those practices. We see how Goodman's archaeologists, in learning to see and represent their seeing, were *becoming* archeologists, and could rely on one another as capable observers. They not only acquire knowledge but an identity and a way of understanding the world, and a basis for relying on one another. Community implies similarity.

If so, how do we fold the work of people from different epistemic communities into the complex assemblages that are digital libraries? One strength of digital libraries like CalFlora is the diversity of representations that they bring together, contributed by people with different training, from different situations, over a prolonged period of time. Furthermore, digital libraries like CalFlora depend on the work of diverse groups to design, implement, manage, and operate them. How can digital libraries operate at this nexus of diversity?

5. ONE DIGITAL LIBRARY'S EXPERIENCE

CalFlora takes observations from a variety of sources and tries to make them useful to an equally diverse set of users. In this section, we look at some of the practical choices that are made about the metadata, rooted in the practices of biodiversity work, that help users understand CalFlora's contents, with special attention to contributions from amateurs. It is not the purpose of this paper to present some ideal solution, but to examine the practices in a field and the practical solutions that one group has developed in light of the discussion about epistemic cultures and communities of practice.

CalFlora is by policy inclusive in its contents. Like [28], suggesting provenance rather than centralization as a safeguard against deception, CalFlora's policy is to be inclusive and devote resources to provenance.

Users tell us that they evaluate observations based on factors internal to the observation, an on the observer. Observations can be evaluated by their plausibility – e.g., an expected taxon in

expected place – and internal consistency, e.g., the reported location fits other indicators of place within the record. However, since observations of *unexpected* occurrences may be particular significant, relying on expectations may eliminate useful information.

For example, one respondent took us through her criteria for plant identifications. A record of a sighting is less credible than a specimen in hand. A report of a taxon in an unexpected place is less plausible than in a place where it is common. She knows some individuals or classes of people (e.g., park rangers) as experts in geographical areas; she'll trust their identification of taxa common to their area, but not necessarily of rare ones. Others are experts on a taxon, on which she'll trust them wherever it may occur.

Discussions within CalFlora identified three factors determining an observer's credibility: the skill of the observer, the observer's relationship to that which is observed (e.g., expertise in a particular taxon or geographical area), and his or her certainty in making this particular identification.

5.1 Photos

Photos come from both institutional and personal sources. Individuals register and are screened by CalFlora staff for minimum skills in photography and plant identification. They are asked some basic questions about themselves and their photography (e.g., experience, equipment).

Photographers provide metadata for their photos, typically taxon identification, date and location of the photo, institutional source (if any), and a link to occurrence records for the same taxon. Photographer's name and a simple biography with their email address and web page, when applicable, is linked to each of their photos.

As we've said, identification errors are sometimes a problem with photos. Records are occasionally reviewed by CalFlora-sanctioned experts; every record is labeled as to whether the taxon identification has been reviewed. Searches can be limit to records whose identifications have been verified.

In addition, any users can register to add annotations, including identity changes. Annotators provide their contact information and credentials or background, and are assigned permissions levels depending on their expertise. Annotations may include a variety of comments on the record, including identification corrections, usually with explanation. Proposed name changes are entered directly or reviewed by CalFlora staff, depending on the annotator's permissions level. Subsequent viewers see the annotation and the identity of the annotator.

5.2 Occurrence Records

Occurrence records have until now come from institutional sources. (An institutional source does not guarantee an institutional or expert observer; historically, many significant collections have relied on amateur collectors and observers [42].)

The CalFlora staff work with data providers to create standardized records from existing datasets. Occurrence records consist of ID number, taxon name, institutional source, date, observer name, location, plus observation type and documentation type (specimen in public museum, documented by voucher or expert confirmation, undocumented report, or literature range description). Not all records contain complete data. Searches return a summary table of matching observations and a map showing the taxon's distribution. Each entry in the table is linked to a complete record with more detail (if available) and to added information about the dataset and the institution, including a contact person.

CalFlora users who wish to supply updates or corrections are advised to contact the institutional source directly. An annotation function, similar to the one for photos (described below), is anticipated.

A process by which individuals may register and contribute plant observations is under development. CalFlora wrestled with how to define and code contributors' skills. As one person said, "You can't just ask people how competent they are." One participant insisted that nothing was needed but the observers' name. She drew the parallel with the herbarium where she works, where the records created over many years always contain the name of the observer, and, she contended, users know which observers are reliable. With a limited set of contributors and users, this was perhaps sufficient. But others in CalFlora felt this was insufficient.

In designing the registration form for contributions, the committee found that identifying professionals and people with no particular qualifications was not hard; the problem was "expert amateurs," people with expertise but no professional qualifications or training. Everyone knew what the phrase meant; the problem was operationalizing it so that contributors could reliably classify themselves. The draft registration form that resulted asks for contact information, "bio/credentials" (a free text field), expertise and interests, institutional affiliation (if any), and experience level. The choices for this last: professional botanist/field biologist, experienced in plant identification and/or regional flora, adult age 18 or over (i.e., an adult with none of the above credentials), teen age 13 to 18, or child under 12.

Registrants are asked to read and abide by two statements: an agreement to submit only one's own first-hand observations, and a quality commitment to use correct scientific names and to "submit uncertain identifications only if I believe them to be very important and time sensitive, and will label such reports 'uncertain'."

Another topic that was debated and eventually dropped was whether to limit use or require registration of users. Throughout its life, CalFlora has been available over the Internet without charge. Some CalFlora participants have complained that government and privately-sponsored biodiversity databases that operate on a cost recovery basis charge such high fees that they are available only to developers, not to people seeking to protect the environment. So CalFlora is committed to being free of charge and freely available.

However, some fear that making CalFlora photos and locations for rare taxa openly available may result in their over-collection or destruction. A lengthy discussion ensued about limiting and/or vetting users, to block, or at least discourage, people who might misuse the data. The committee could not agree on the need and a mechanism for limiting use. The proposed solution for protecting sensitive data has been to fudge locations in whatever way is requested by the data owners, and for CalFlora to seek out information on species affected by vandalism and illegal collecting and to review and decide upon requests for suppression of location information for specific taxa.

6. DISCUSSION

I contend that the lessons from CalFlora apply to other kinds of digital libraries and information systems, particularly those that allow the sharing of "unpublished" data. Knowledge management systems that access internal documents, data warehousing systems that combine data from multiple sources over time, and other digital libraries systems that provide access to data from multiple sources of varied quality all have to address the trustworthiness of contents and sources and appropriate use.

From the discussion of epistemic cultures and communities of practice, we take, first of all, an emphasis on the *social nature of knowledge*, our dependence on others, and our need to decide whom and what we believe.

Second is an emphasis on *practice*, the actual, day-to-day activities of work by which people perform their work and make it accountable.

Third is *difference*. Different epistemic communities have different epistemic mechanisms, technical and social: different practices of work, of determining what is true or credible, and who is trustworthy. While these mechanisms may be largely continuous across fields, especially within science, there are also differences, which may be subtle. The similarities among fields may even mask their subtle differences, as Knorr [20]shows that two laboratory-based sciences have very different concepts of the laboratory.

Fourth, is the idea that *knowers are produced by epistemic cultures and communities of practice*. People's understandings of the world, of themselves, of what they are doing, what is important, and what is valuable are a result of the community in which they are trained and in which they participate. The process of learning is one of *becoming*, of taking on an identity.

In CalFlora, we see a commitment to provide users with detailed information to help them assess the credibility of both the observation and its source, as well as to make the observation data maximally useful. First, detailed metadata about each observation helps the user to understand the data and make use of it according to his or her own practices. For example, one element in occurrence records is observation type: different methods of observation are biased in favor of or against rare taxa. An informed user will take this into account. Methods of identification differ in their reliability; for some purposes, a user will want only the most reliable identifications, and for other purposes, less so.

Second, CalFlora works closely with data providers to ensure that they are comfortable with how their data are presented. Participants exercise considerable control, accommodating differences in data providers' understandings and standards.

Third, CalFlora provides a variety of clues about the training and practices of the observer. CalFlora insists on contact information and descriptions of credentials or expertise from contributors. The free text qualifications fields may be more useful than a closed-ended set of categories: what people choose to say about themselves and how they say it can be revealing. In addition, the observation records themselves may provide clues to the observer's expertise. Someone who can provide the scientific name of a taxon and latitude and longitude is probably not a casual hiker.

CalFlora's policies are aimed at providing maximum flexibility in the use of data, and maximum autonomy for the data user. (CalFlora has defaults for functions such as mapping so that users can rely on the expertise of CalFlora if they wish.) Flexibility and autonomy are necessary to accommodate multiple knowledges; maximum information is needed for assessing similarity of epistemic machineries. Our point that not that CalFlora is designed with epistemic cultures and communities of practice in mind. Our point is that the way that CalFlora's decision-makers have designed it to be useful provides us with insight into the workings of epistemic communities. The perception of CalFlora's decision-makers that users need this deegree of flexibility and autonomy is consistent with notions of epistemic cultures and communities of practice. In deciding what and whom to trust, people look for others who share their methods and understandings; and what they consider adequate evidence to justify credence will vary.

6.1 Digital Library Design

What are the implications of this discussion for digital libraries other than CalFlora? DLs need to be designed to suit the practices of the specific epistemic groups it is intended to serve. It must provide maximum information about the provenance of its contents and about its contributors, balancing privacy with the users' need to know. Since we trust the most members of our same (or a closely-related) epistemic community, the digital library needs to facilitate the processes by which people make these assessments. Such indicators as training and experience and institutional affiliation are useful, but members of an epistemic community use a range of evidence. CalFlora's free text fields for bio/credentials and expertise are apt; let contributors decide what they need to say, and let users judge.

CalFlora also provides its users with maximum flexibility in searching, retrieving, downloading, and re-using its contents. For users to be able to accommodate their own practices, and not just those that the digital library designers built in, maximum freedom in using the contents is necessary. This is possible in CalFlora because contributors agree to it; in digital libraries with other ownership arrangements, this can be a problem.

Users also need to know the "provenance" of the digital library itself: who designed it, and for what purpose? What design choices were made? Including, what is *not* visible to the user? (For example, in CalFlora, how precise are the locations?)

Another implication is added evidence for the importance of user-centered design. Involving users in design is crucial, since only they understand the complexities of their epistemic machineries.

However, this does not necessarily imply highly-customized, fragmented DLs. Elsewhere [47], I have discussed digital libraries as boundary objects [41, 42], which are both plastic enough to adapt to local needs and have different specific identities in different communities, and robust enough to maintain a common identity across sites, and be a locus of shared work. The power of CalFlora is in its ability to bring together multiple groups and data sources; attention to the differences across epistemic cultures has to be balanced by their need to share information, which is the role of boundary objects.

Differences across epistemic groups help to explain why the process of involving users in design is often much more difficult than one might expect. Most digital libraries serve a variety of user groups, whose deeply-embedded, differing knowledges complicate the design process. Groups are likely to differ, not only in what they want of a digital library, but in their assumptions about knowledge and the knowledge work that the digital library is designed to facilitate, about what is known and the methods and standards for determining what is knowledge and who is competent to speak.

6.2 Digital Library Processes

These observations also have implications for the design process and the management and operation for digital libraries. It is not only users who are members of different epistemic communities. Most digital libraries are the product of multi-disciplinary collaborations that may include technical specialists, representatives of user groups, and (sometimes) librarians.

CalFlora is designed, operated, and funded by a coalition of groups. Differences in epistemic machineries, in understandings of the world, and in values surface in CalFlora's decision-making processes. The result is an on-going need for negotiation among participants. Elsewhere [47] I describe DLs as actor-networks. Translation and enrollment are never finished.

Suchman's [43] reflections on the construction of technological systems are useful here. Also writing from a perspective of multiple knowledges and varied epistemic communities, she sees the development of technical systems as entry into a network of working relations between designers and users that make technical systems possible, rather than as the creation of discrete devices.

She reflects on her experiences working across the boundaries between users and designers. Crossing boundaries, she says, means "encountering difference, entering onto territory in which we are unfamiliar and, to some extent, therefore, unqualified" – and uncomfortable (p. 25). Useful system design requires the on-going creation of situations for "the meeting of different partial knowledges" (p. 25)...in "an increasingly dense and differentiated layering of people and activities, each operating within a limited sphere of knowing and acting that includes variously crude or sophisticated conceptualizations of the others" (p. 30).

In summary, networked information systems are the loci of a web of relations among different epistemic communities, including varied user groups, technologists, and others engaged in digital library design and management. Wherever people are working at the boundaries of knowledge communities we will see differences and disconnects, negotiations and assessments. The boundary crossing work of designing and building systems, populating them with content, and keeping them operational requires work at comfortable boundaries, and the exercise of mutual trust.

7. ACKNOWLEDGEMENTS

Many thanks to the staff of CalFlora, Ann Dennis and Tony Morosco, and the CalFlora Advisory Committee for giving me access to their deliberations, and to Ann and Tony for helpful comments on an earlier draft.

8. REFERENCES

[1] Alexander, J., Tate, M. *Web Wisdom: How to Evaluate and Create Information Quality on the Web*. Lawrence Erlbaum Associates: Mahwah, NJ, 1999.

[2] Bishop, A. P., Buttenfield, B., Van House, N. A. (eds.). *Digital Library Use: Social Practice in Design and Evaluation*. MIT Press: Cambridge, MA, 2001.

[3] Bowker, G. C. Biodiversity datadiversity. Social *Studies of Science 30*, 5 (2000) 643-683.

[4] Burbules, N. C. Paradoxes of the web: the ethical dimensions of credibility. *Library Trends 49*, 3 (2001) 441-453.

[5] Davenport, E., Cronin, B. The citation network as a prototype for representing trust in virtual environments. In Cronin B., Atkins H. B. (eds.). *The Web of Knowledge: a Festschrift in Honor of Eugene Garfield*. Information Today Inc. & The American Society for Information Science: Medford, NJ, 2000.

[6] Donath, J. Being real: questions of tele-identity. In Goldberg K. (ed.). *The Robot in the Garden: Telerobotics and Telepistemology in the Age of the Internet*. MIT Press: Cambridge, MA, 2000.

[7] Fogg, B. J. and others. What makes web sites credible? a report on a large quantitative study, *in CHI 2001* (Seattle, WA, 2001), ACM, 61-68.

[8] Fritch, J. W. and Cromwell, R. L. Evaluating internet resources: identity, affiliation, and cognitive authority in a networked world. *Journal of the American Society for Information Science and Technology 52*, 6 (2001) 499-507.

[9] Fukuyama, F. *Trust: the Social Virtues and the Creation of Prosperity.* Free Press Paperbacks: New York, 1995.

[10] Fuller, S. *Social Epistemology*. Indiana University Press: Bloomington and Indianapolis, IN, 1988.

[11] Gambetta, D. Can we trust trust? In Gambetta D. (ed*.). Trust: Making and Breaking Cooperative Relationships*. Basil Blackwell: New York, NY, 1988.

[12] Gambetta, D. *Trust: Making and Breaking Cooperative Relationships.* Basil Blackwell: New York, NY, 1988.

[13] Goldman, A. I. *Knowledge in a Social World*. Clarendon Press: Oxford, 1999.

[14] Goldman, A. I. Social epistemology. *Stanford Encyclopedia of Philosophy* . 2001. 10-19-2001.

[15] Goodwin, C. Professional vision. *American Anthropologist 9*6, 3 (1994) 606-634.

[16] Greenspan, S., Goldberg, D., Weimer, D., and Basso, A. Interpersonal trust and common ground in electronically mediated communication, in *CSCW '00* (Philadelphia, PA, 2000), ACM, 251-260.

[17] Hess, D. *Science Studies: an Advanced Introduction*. New York University Press: New York, 1997.

[18] Jensen, C., Farnhman, S. D., Drucker, S. M., and Kollock, P. The effect of communication modality on cooperation in online environments, in *CHI 2000* (The Hague, Amsterdam, 2000), ACM, 470-477.

[19] Jones, K. Trust. *Routledge Encyclopedia of Philosophy*. 2000. 7-18-2001.

[20] Knorr Cetina, K. *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard University Press: Cambridge, MA, 1999.

[21] Lane, C. Introduction: theories and issues in the study of trust. In Lane C., Bachman R. (eds.). *Trust Within and Between Organizations: Conceptual Issues and Empirical Applications.* Oxford University Press: Oxford, 1998.

[22] Latour, B. Drawing things together. In Lynch M., Woolgar S. (eds.). *Representation in Scientific Practice.* MIT Press: Cambridge, MA, 1990.

[23] Latour, B., Woolgar, S. *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press: Princeton, NJ, 1991.

[24] Lave, J. The practice of learning. In Chaiklin, Seth, Lave J. (ed.). *Understanding Practice: Perspectives on Activity and Context*. Cambridge University Press: Cambridge, 1983.

[25] Lave, J. *Cognition in Practice: Mind, Mathematics, and Culture in Everyday Life*. Cambridge University Press: Cambridge, 1988.

[26] Lave, J., Wenger, E. *Situated Learning: Legitimate Peripheral Particiption.* Cambridge University Press: Cambridge, England, 1991.

[27] Luhmann, N. *Risk: a Sociological Theory*. Aldine de Gruyter: New York, 1994.

[28] Lynch, C. A. When documents deceive: trust and provenance as new factors for information retrieval into tangled web. *Journal of the American Society for Information Science and Technology* 52, 1 (2001) 12-17.

[29] Lynch, M., Woolgar, S. (eds.). *Representation in Scientific Practice.* MIT Press: Cambridge, MA, 1990.

[30] Maier, D., Landis, E., Cushing, J., Frondorf, A., Silberschatz, A., and Schnase, J. L. Research Directions in Biodiversity and Ecosystem Informatics: Report of a NSF, USGS, NASA Workshop on Biodiversity and Ecosystem Informatics. Greenbelt, MD, NASA Goddard Space Flight Center, 2001.

[31] McDonald, D. W. and Ackerman, M. S. Expertise recommender: a flexible recommendation system and architecture, in *CSCW '00* (Philadelphia, PA, 2000), ACM, 231-240.

[32] Olson, G. M. and Olson, J. S. Distance matters. *Human-Computer Interaction 15* (2000) 139-178.

[33] Olson, J. S. and Olson, G. M. i2i trust in e-commerce. *CACM 43*, 12 (2000) 41-44.

[34] Putnam, R. D. *Bowling Alone: the Collapse and Revival of American Community*. Simon & Schuster: New York, 2000.

[35] Rieh, S. Y. and Belkin, N. J. Understanding judgment of information quality and cognitive authority in the WWW, in *American Society for Information Science and Technology Annual Meeting*, Pittsburgh, Pennsylvania, Oct. 24-29, 1998, 1998), Information Today, Inc, 279-289.

[36] Schiff, L., Van House, N. A., and Butler, M. Understanding complex information environments: a social analysis of watershed planning, in *Digital Libraries '97: Proceedings of the ACM Digital Libraries Conferenc*e (Philadelphia, PA, 1997), ACM Press, 161-186.

[37] Schneider, F. B. ed. *Trust in Cyber*space. National Academy Press: Washington, DC, 1999.

[38] Seligman, A. B. *The Problem of Trust*. Princeton University Press: Princeton, NJ, 1997.

[39] Shapin, S. *A Social History of Truth: Civility and Science in Seventeenth-Century England*. University of Chicago Press: Chicago, IL, 1994.

[40] Shneiderman, B. Designing trust into online experiences. *CACM 43* (2000) 57-59.

[41] Star, S. L. The structure of ill-structured solutions: boundary objects and heterogeneous distributed problem solving. In Gasser, L., Huhns, M. (eds.). *Distributed Artificial Intelligence*, 2. Pitman Publishing: 1989.

[42] Star, S. L. and Griesmer, J. R. Institutional ecology, "translations," and boundary objects: amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Scienc*e *19* (1989) 387-420.

[43] Suchman, L. Working relations of technology production and use. *Computer Supported Cooperative Work (CSCW) 2* (1994) 21-39.

[44] Sztompka, P. *Trust: a Sociological Theory.* Cambridge University Press: Cambridge, England, 1999.

[45] Uslaner, E. M. Trust, Civic Engagement, and the Internet. 2000.http://www.pewtrusts.com/pdf/vf_pew_internet_trust_paper.pdf.

[46] Van House, N. User needs assessment and evaluation for the UC Berkeley electronic environmental library project, *in Digital Libraries '95: The Second International Conference on the Theory and Practice of Digital Libraries* (San Antonio, TX, 1995).

[47] Van House, N. A. Digital libraries and collaborative knowledge construction. In Bishop A. P., Buttenfield, B., Van House, N. A. (eds.). *Digital Library Use: Social Practice in Design and Evaluation.* MIT Press: Cambridge, MA, 2002.

[48] Van House, N. A. Digital libraries and practices of trust: networked biodiversity information. *Social Epistemology*, in press (2002) .

[49] Van House, N., Butler, M., and Schiff, L. Cooperative knowledge work and practices of trust: sharing environmental planning data sets, in *CSCW '98: The ACM Conference On Computer Supported Cooperative Work* (Seattle, WA, 1998), 335-343.

[50] Vann, K. and Bowker, G. C. Instrumentalizing the truth of practice. *Social Epistemology* 15, 3 (2001) 247-262.

[51] Wenger, E. *Communities of Practice: Learning, Meaning, and Identity.* Cambridge University Press: New York, 1998.

[52] Wilson, P. *Second-Hand Knowledge: an Inquiry into Cognitive Authority*. Greenwood Press: Westport, CT, 1983.