SOCIAL EPISTEMOLOGY, 2002, VOL. 16, NO. 1, 99–114

Digital libraries and practices of trust: networked biodiversity information

NANCY A. VAN HOUSE

Egan and Shera's (1952) groundbreaking paper has three main thrusts. One is their argument for social epistemology as the disciplinary framework for understanding the intellectual processes of society. The second is the importance of 'graphic communication' and bibliography in these processes. The third is the need for an empirical methodology for studying social epistemology and graphical communication. This paper presents findings from an empirical study of social epistemology and graphical communication in a world of networked information. The Internet and digital libraries make it possible to share unpublished data more readily and more widely than ever before, giving rise to new problems in bibliography and graphic communication. This paper looks specifically at issues of credibility and trust in the sharing of biodiversity data.

Egan and Shera identify a fundamental problem in their day as a conflict between what they term 'macroscopic' and 'microscopic' approaches to bibliography. They say that then-current methods of bibliography were microscopic, concerned with the bibliographic needs of small groups with common interests. They argue for a more macroscopic approach, addressing bibliographic communication among groups of scholars and between groups at various levels, including scholars, practitioners, educators and the lay public, and the co-ordination of both the various types of bibliographic services and of groups themselves.

Methodologically, Egan and Shera argue for:

> an objective analysis of the several types of communication used by the [researcher] in the prosecution of his [sic] work ... an examination of the bibliographic devices best suited to expedite the flow of each type of communication, and ... the means by which the results of his work are made known not only to his colleagues but to other individuals or groups who might be interested. (1952, p. 128)

They recommend 'situational analysis', the study of the 'kinds of information, knowledge, and insights' needed in various situations. Although the number of such situations is great, and each unique, they argue that situations probably fall into a finite number of types. The goal would be 'not to supply final answers concerning the informational requirements of specific groups engaged in specific types of

Author: Author: Nancy A. Van House, School of Information Management and Systems, University of California, Berkeley 94720-4600, USA; e-mail: vanhouse@sims.berkeley.edu

Social Epistemology ISSN 0269-1728 print/ISSN 1464-5297 on line © 2002 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/02691720210132833

activities but to develop a *sound methodology* by means of which situational analysis can be applied ... to a variety of differing situations or to the same situation as it changes over time' (pp. 136–5; original emphasis).

Networked information has given a new dimension to Egan and Shera's micro/macro dilemma. Networking changes the way that work extends across space and time. At the micro level, digital information is more readily customized to the needs and practices of specific communities. At the macro level, digital graphic records are more mobile, more easily shared across groups but also more easily removed from the context of production that helps the user to understand them.

This paper presents insights from a situational analysis of a group of people engaged in biodiversity work and in the sharing of data via a publicly accessible, web-based digital library. The field of biodiversity research is heavily engaged in the creation of heterogeneous sets of digital information, creating new problems in the flow of graphic communication and engendering new solutions, as well as making more visible old problems and solutions. My purpose is to examine issues of data sharing and, in particular, credibility of networked information in a multi-disciplinary, often-political field.

The data come from an ethnographically informed qualitative study. This paper will focus on the findings rather than the method (which is described further in Van House, 2002). But I argue that ethnographic fieldwork, which is being used increasingly in information systems and digital library work (Van House *et al.*, 2002) is an appropriate response—perhaps the most appropriate response—to Egan and Shera's call for the development of methods for situational analysis.

Over the last several years, my colleagues and I have conducted interviews with people engaged in environmental work in the interests of learning more about their information practices and their uses and concerns about digital libraries (Van House, 1995; Schiff *et al.*, 1997; Van House *et al.*, 1998; Van House, 2002). This paper will draw primarily on interviews with people associated with CalFlora (http://www.calflora.org), a digital library of plant observation records and images.

My primary argument is that networking allows information to cross social and technical boundaries that have provided the context for assessing its credibility and meaning, throwing into relief practices of trust. Sharing information requires that users and providers trust one another. Networked information raises questions at each end of the information transaction. Not only do users have to assess the credibility of information, but providers are concerned about possible misuse of their data. This paper addresses these issues in the context of the notions of communities of practice (Lave, 1988; Lave and Wenger, 1991; Wenger, 1998) and epistemic cultures (Knorr Cetina, 1999).

My interest is not so much in the specifics of CalFlora as in how changes in information technology and organizations foreground the taken-for-granted practices by which groups solve problems of social epistemology on a day-to-day basis; in this case, problems associated with the sharing of data. Digital libraries give us an opportunity to study these processes as they are made more visible.

Digital libraries, and, by extension, other kinds of information systems, must be designed to support existing and emergent practices that respond to the ease with which digital information moves within and across groups of specialists and non-specialists. The process of instantiating into code the information needed for assessing credibility requires a conscious reflection and negotiation on issues and practices that are often invisible, and may differ across epistemic communities.

1. *Biodiversity data and networking*

Biodiversity research, which is concerned with the diversity of life and the ecosystems that maintain it, has extremely complex data needs, including the creation and convergence of sets of heterogeneous databases (Bowker, 2000).

> [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. (Maier *et al.*, 2001, p. 3).

An important source of fine-grained biodiversity data is the observer in the field. Disciplines such as botany, ornithology and astronomy have long relied on 'citizen science', the observations of amateurs with expertise but no formal credentials. To use their work requires systems of recording, collecting and collating their observations and assessing the accuracy of specific reports: e.g. is the report of a rare species in an unusual location a significant sighting or an identification error?

This reliance on amateur observers is not new. 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 reports from the public over the web on an on-going basis.

The Internet makes more, and more varied, information available from a growing range of formal and informal sources. In earlier research, my colleagues and I (Van House, 1995; Schiff *et al.*, 1997; Van House *et al.*, 1998; Van House, 2002) investigated the concerns of both data users and producers about networked information. The question for users is that of testimony: how credible are these data? How is the user to understand them? What situational factors affect their interpretation and use? Networked information raises the profile of such concerns because data may by-pass the traditional institutional authorities, such as the publication system, by which information has been warranted and by which users establish not only information's reliability but also its meaning. For users, the result is a greater responsibility to interpret and assess the validity and credibility of information from a growing range of familiar and unfamiliar sources.

Data producers reported increasing demands to make their 'raw' data usable, not only to their colleagues, but also more widely via the Internet. The people we talked with had shared, before the Internet, data via personal contacts or within closed professional communities. The Internet depersonalized these processes and increased the number and variety of people using their data.

Data owners reported a variety of concerns (Van House, 2002). The one most relevant to this discussion is fear of misuse. Misuse, in respondents' terms, could result from two main causes. The first is lack of competence (as defined by data owner). For example, estimating population parameters from sample data, differentiating trends from random fluctuations and combining data from different

sources compiled in different ways at different times all require an understanding of the data and the underlying science and expertise in its use.

Another kind of misuse identified by our respondents refers to the purposes for which data are used. Environmental planning and biodiversity work is highly political, with potentially great environmental and economic consequences. Some respondents were much concerned that their work could be used either to harm the environment or to undercut their own analyses and conclusions. Most recently, the California Department of Water Resources has asked the UC Berkeley Digital Library Project, a publicly accessible web-based digital library, to remove photos and maps of components of the state's water system that might be subject to attack by terrorists.

2. *The case study*

CalFlora[1] (http://www.calflora.org) is a non-profit organization that supports a comprehensive web-accessible database of plant distribution information for California. It provides ready access to data needed to identify critical issues in conservation of plant diversity and to analyse consequences of land use alternatives and environmental change on the distribution of native and exotic species. It serves researchers and the general public.

For this analysis, I am interested in two components of CalFlora. One is a database of over 800,000 reports of observations of plants in California—specimens in collections and reports from the field—from over a dozen government and private agencies. The other is a database containing over 22,000 photos of California plants, which is a joint product of CalFlora and the UC Berkeley Digital Library Project (http://elib.cs.berkeley.edu). These photos come from a variety of institutional and personal sources. Contributors retain ownership over their data, and can withdraw it at any time.

In this paper, I consider CalFlora's policy debates in two areas. Currently, photographic contributions are accepted from the public; CalFlora is developing policies and procedures for accepting occurrence observations from individuals, as well. The other is the inclusion of locations and images of specimens of endangered taxa. The debate hinges on whether this information would encourage their preservation or lead to their destruction.

3. *Epistemological trust*

If knowledge is social, and knowledge work entails the distribution of cognitive labour, then people have to decide whose work to accept—the problem of testimony.

> 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 moral relation is *trust* ... I argue that the identification of trustworthy agents is necessary to the constitution of any body of knowledge ... [W]hat we call 'social knowledge' and 'natural knowledge' are hybrid entities: what 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. (Shapin, 1994, original emphasis)

In the world of networked information, more ready access to information means more difficulty in deciding whom and what we will rely upon, and how we will decide whom

to trust. In this paper, I examine the role of the knowledge community in the identification of trustworthy others. In this section, I examine the issue of trust, specifically epistemological trust. Then I look at the role of epistemic communities in determining trust. I look at how the boundaries of epistemic communities are drawn and crossed. I then relate epistemic trust and communities to some of the issues CalFlora is currently confronting.

Trust is a topic in, among other areas, philosophy (Jones, 2000), sociology (Luhmann, 1994; Sztompka, 1999) and political science (Gambetta, 1988b). One approach (e.g. Gambetta, 1988a) focuses on calculative trust (Lane, 1998): the assessment of risk, predictions of others' behaviour and rational choice. Generally, these discussions are based in economics and political science, and concerned with the division of labour, contracts and exchanges.

A second approach is concerned with the role of trust in the social order, the relationship of trust to citizenship, co-operation, and morality (Fukuyama, 1995; Seligman, 1997; Sztompka, 1999; Putnam, 2000). It stresses reciprocity and duty toward the community based on habit and moral obligation rather than rational calculation or sanctions, and a social order based on voluntary co-operation.

The collective nature of knowledge foregrounds a third type of trust, which has been called epistemological trust (Davenport and Cronin, 2000) or the granting of epistemic authority (Goldman, 1999). Shapin (1994) reviews the literature on trust and concludes that, while the role of trust in the order of society has received a lot of attention, 'the role of trust and authority in the constitution and maintenance of systems of valued *knowledge* has been practically invisible' (p. 16; original emphasis). This paper is intended to help increase the visibility of the role and especially the practices of trust in systems of knowledge and bibliography, specifically networked data.

Accepting others' testimony is, among other things, a strategy of cognitive efficiency. We have neither the ability nor the resources to test all possible knowledge claims. Nor do we necessarily wish to; only some knowledge claims are of sufficient importance to us for us to engage in detailed examination. Trust reduces transaction costs, in this case, the costs of verifying knowledge claims on our own. Furthermore, Wilson (1983) points out that we generally do not evaluate many claims; we wait until we need to decide whom or what to believe, and even then we may weigh the costs of evaluating claims against the penalties of believing wrongly.

4. *Trust and networked information*

Networked information gives people more ready access to more sources of information, increasing both the amount of information available and the variety of unfamiliar sources to be assessed. Discussions about trust in the networking world falls into four categories. The first is trustworthy systems, which do what people expect despite environmental disruption, human user and operator errors, and hostile attacks (Schneider, 1999). The second is contract-like online agreements, e-commerce and e-services (Shneiderman, 2000). A third area, covering virtual communities and online social interactions, is concerned with questions about the presentation of self and identity and the possibilities for deception, hurt feelings and embarrassment (Donath, 2000; Friedman *et al.*, 2000).

Of particular relevance to this paper is a fourth area, concerned with the credibility of web resources, evaluative criteria for web sources and the extent to which people use

questionable information from the web (Rieh and Belkin, 1998; Alexander and Tate, 1999; Fritch and Cromwell, 2001)—that is, epistemological trust. Some of this work is descriptive; other is normative.

A major criterion that people use to assess web-based information is the credentials of the author and/or the sponsoring institution. Burbules (2001) reviews issues and methods of determining the credibility of online materials and 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). Burbules 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).

In sum, then, evaluation of information on the Internet relies heavily on our evaluation of the source, and on the judgments of the community in which we have confidence. The questions then are, how do we evaluate sources, and how do we choose the community whose judgments we trust?

5. *Assessing credibility*

As the quote from Shapin above claims, assessments of credibility depend largely on our assessments of individuals. Even Giddens (1990), who identifies one characteristic of modernity as reliance on expert systems of knowledge instead of known others, argues that trustworthy others remain important. Giddens speaks of 'access points', the people who are our contacts with expert systems, such as doctors and the medical system, as the place where face-to-face relations and expert systems come together. Similarly, Shapin ends his study of seventeenth century English experimental science with a reflection on contemporary science and concludes that 'trust in known persons' (p. 415) remains a key element in scientific and other forms of knowledge.

What are the criteria for determining credible people? The answers often given are competence and honesty (e.g. Wilson, 1983; Goldman, 1994). Based on my empirical work, I would add, at least in the domain of observational data in biodiversity, shared orientations and values, or, to give it a shorthand label, 'virtue'.

Competence is relative; we recognize degrees and spheres of competence (Wilson, 1983). In biodiversity, competence is assessed in both the methods and the content of the work. In assessing observational data, the user asks, did the observer use appropriate methods for reporting and collecting data? And does she or he have the expertise to, for example, make accurate taxon identifications?

When we cannot evaluate competence based on personal knowledge or performance, we rely on such indicators as occupational or educational credentials and reputation among people whom we judge to be knowledgeable (Wilson, 1983), and other indicators in areas that we can assess. Mechanic and Meyer (2000) found that patients assess their doctors based on interpersonal competence, that is, caring, concern and compassion; the authors concluded that, while patients stated a strong interest in technical competence, they could not assess that; they could however assess interpersonal competence.

Honesty is the propensity to tell the truth. Deception can take many forms for many purposes. Lynch (2001) reports a new form of deception in the networked world: deceiving search engines to increase the ranking of a web document among retrieved results. Lynch 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 implement this: a centralized, formalized approach to deciding what is included 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*.

There is more to being a reliable source of information than the ability to do good quality work and a lack of deception, however. In our discussions with people engaged in biodiversity work, it was clear that in their world there are good deeds, preserving the natural environment, and bad deeds, destroying it. For example, the current CalFlora debate about whether to provide locations for occurrences of endangered taxa hinges on the threat posed by people who would over-collect horticulturally desirable taxa or destroy rare populations—for example, landowners seeking to avoid legal constraints on the use of their land. Some of the people we talked to in water planning assumed that the resource extraction industries (such as timber harvesting) would hide or slant data to ensure that it reflected well on their activities. For this we adopt, from Shapin's discussion of gentlemanly behaviour in 17th-century England, with tongue slightly in cheek, the term 'virtue'. This is something more than the value- or norm-based trust identified by sources such as Lane (1998). Shared orientation and values, or at least acceptable orientation and values, were important in relation to misuse of data as well as untrustworthy sources.

The qualities of the originator of information, then, are critical in assessing its credibility. This can be termed provenance: from where and from whom does the information come, and what do we know about the originators or guarantors? The question is not only whether we consider them competent, and honest, but 'virtuous'—whether they embody our values and view of the world and our standards for work.

6. *Epistemic communities, practice, and culture*

Education, training, experience, reputation and many other criteria by which we assess credibility can be seen as indicators of membership in a profession, discipline or knowledge community. The interdependence of community, practice, culture and knowledge has been an important topic in contemporary social theory and science studies. Many strands of contemporary social theory emphasize that knowledge depends on local contexts and activity: 'knowledge … does not stand outside of practical activity: it is made and sustained through situated practical activity' (Shapin, 1994, p. xix).

Two areas of investigation particularly interested in knowledge are science studies and learning. Science studies in the 1970s and 1980s became interested in scientific practice as what scientists actually do, as well as the material resources of their work. Although the concept of practice in science studies has taken many forms (Pickering,

1992), they have in common a focus on agency, activity and the social in generating knowledge.

Lave and Wenger (Lave, 1988; Lave and Wenger, 1991; Wenger, 1998) develop a practice-based approach to learning that contextualizes knowledge and legitimates different knowledges (Vann and Bowker, 2001). They develop the notion of 'community of practice' to emphasize the mutuality of community, knowledge, activity and social practice. Learning is not the absorption of pre-given knowledge but a creative act of the whole person acting in the world, It is not located in the mind of the individual, but in the relations among practitioners, practice, artefacts and the social organization of communities of practice. Communities of practice imply

> participation in an activity system about which participants share understandings concerning what they are doing and what that means in their lives and for their communities ... A community of practice is an intrinsic condition for the existence of knowledge, not least because it provides the interpretive support necessary for making sense of its heritage. Thus, participation in the cultural practice in which any knowledge exists is an epistemological principle of learning. (Lave and Wenger, 1991, p. 98)

To this emphasis on practice and communities in knowledge we can add, from science studies, Knorr Cetina's (1999) notion of epistemic cultures, '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' (p. 1; original emphasis).

Knorr Cetina argues that epistemic cultures are structural features of knowledge societies, including but not limited to science. While praising the emphasis on practice in science studies, she claims that there has still been too little attention to the machineries of knowing, their depth, and particularly their variety in different epistemological settings. They differ, she says, not just between science and non-science, but even within science. Her book consists of empirical investigations and comparisons of the differing epistemic cultures of experimental high-energy physics and molecular biology.

In sum, the emphasis on practice, communities and cultures in some threads of contemporary social theory and science studies implies that knowledge, understanding and identity are intimately tied to practice in community. Cultures are the amalgams of practices and mechanisms that determine how and what we know, how we see the world, and our identities and values. And epistemic cultures differ, even within science. These differences have implications not only for how we understand epistemic communities, but also for how they interact.

Membership in a community of practice, engagement in an epistemic culture, imply similarities of practice and shared understandings and meaning that not only help us to identify people whom we can probably trust, but that shape us as knowing subjects.

7. *Boundaries*

The boundaries of these communities at least help us to determine whom to believe, who's in and who's out. These boundaries are not natural but constructed. And they are continually redrawn.

> To miss the interpretive work that creates contexts for decisions about who to trust with reality is to lose the sociological handle on what is happening. When considered as a cultural space constructed in boundary work, science becomes local and episodic rather than universal; pragmatic and strategic rather than analytic or legislative; contingent rather than principled; constructed rather than essential. (Gieryn, 1999, p. 27)

But boundaries also tend to persist. Giddens (1984) among others argues that structures tend to be produced and reproduced in interaction. In the tension between stability and reproduction, continuity and change, boundaries tend to be self-reinforcing.

The publishing system draws and reflects boundaries of many kinds: between published and unpublished information, among different kinds of information, and among different user groups and knowledge communities. In networked information, the ease with which information may by-pass publication means that these boundaries become much more fluid than ever before. And information users are more often able to, or required to, draw their own boundaries, in areas in which they may or may not have expertise themselves.

8. *Boundary crossings*

Boundaries are drawn, enacted and enforced, but also crossed. Scientific work of all kinds requires co-operation. Texts and other forms of graphic records are a primary method of crossing boundaries.

Biodiversity research requires the convergence of heterogeneous databases across space, time, and disciplines (Bowker, 2000). A number of investigators have asked how scientists manage collective action and stabilize facts across communities or social worlds, however temporary. Actor network theorists (Callon *et al.*, 1986; Latour, 1987; Latour and Woolgar; 1991; Law, 1992: Law and Hassard, 1999) describe scientists as garnering credibility and resources by enrolling human and non-human actors as allies. They give texts and other forms of inscription a primary role in carrying work across space and time and translating interests (Callon *et al.*, 1986).

While actor network theorists emphasize a single actor or viewpoint enrolling allies by translating their multiple interests into one 'obligatory point of passage', Star and Griesemer (Star, 1989; Star and Greisemer, 1989) emphasize simultaneous translations of multiple interests. They note that scientific work requires information that can be used by multiple users and communities for a variety of purposes, retaining its integrity across space, time, and local contingencies without losing its specific meaning in a local setting. They propose that this need is met by what they call 'boundary objects', which are 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. They developed the idea of the boundary object in a study of the creation of the University of California, Berkeley, Museum of Vertebrate Zoology (MVZ), a collection of specimens of amphibians, birds, mammals and reptiles with standardized metadata.

As with the MVZ, to exist, CalFlora has to be of value to an array of participants: contributors, users, and technologists (Van House, 2002). It has to contain credible information in usable form. And the participants have to be in agreement about policies concerning contents, funding, contributions, access and operations. Boundary crossings are always temporary; boundary objects are always threatened with instability.

9. *The case study: policy decisions*

In this section, I will look at two sets of decisions currently facing CalFlora, being made by a group representing CalFlora staff, contributors and users. CalFlora is, by policy, inclusive in the records that it accepts and the users it allows access. The result is a series of tensions and decisions about boundaries and boundary crossings, about what and who is in and out, and how to assess and demonstrate credibility; and an increased burden on users to assess credibility and draw boundaries.

9.1. *Describing contributions*

The photos have come from both institutional and personal sources, including the general public. Most occurrence records have come from institutional sources, but CalFlora is planning to accept occurrence records from individuals as well. One current area of decision-making is how CalFlora can feasibly describe these contributions in ways that will help users assess credibility and what information will be useful and feasible to ask of contributors. The second set of decisions is about the publication of sensitive information that some fear could be misused. The final section of this paper will show how these issues relate to questions of epistemic trust, knowledge communities and boundaries.

The major quality issue with photos and plant occurrence records is taxon identification. Many differences in taxon identifications are possible. Nomenclature changes; what was the appropriate identification when a record was added may be no longer.[2] Outright errors are also a problem. The first photos came largely from one source, a dedicated amateur, who had difficulty with some identifications.

Each photo and occurrence record is linked to the institutional and/or personal source. The purpose is to allow users to credit sources, evaluate observers' credentials and contact data owners about corrections or permissions or for more information. Institutional records include information about the organization and the collection.

Photo contributors register and are screened by CalFlora or UC Berkeley Digital Library Project staff for minimum skill level in photography and plant identification. They are asked basic questions about themselves and their photography (e.g. experience, equipment). A simple biography, with the person's e-mail address and web page, when applicable, is linked to each of their photos. Then they answer questions about each photo as it is added, such as date, location, and (when applicable) plant identification.

The CalFlora decision-making group has been more concerned with assessing observers' credentials for occurrence records than photos. Although not all taxa can be differentiated from photos, a photo provides evidence to substantiate the identification. Furthermore, plant occurrences can be much more significant than photos. An observation may represent the presence of a population in a location that alters our understanding of the taxon's distribution.

Discussions within CalFlora have identified three factors determining credibility of taxon identifications: the skill of the observer, the method of the determination of the identity of the taxon and the observer's certainty in making this particular identification.

A lengthy discussion among the CalFlora staff addressed how to define the skills of contributors. As one person said, 'You can't just ask people how competent they are'. In professional communities, the identity of an observer is often sufficient. One CalFlora participant from a herbarium insisted that nothing was needed but

the observer's name. She drew the parallel with her herbarium, where the paper records created over many years always contain the name of the observer, and, she contended, users know who is reliable. But CalFlora is wrestling with the problem of open access by both users and contributors who do not necessarily know one another.

The CalFlora group concluded that they can only assess skill by asking contributors to describe their own level of expertise. The expert amateur presented the biggest problem, since their expertise rests on knowledge and not training. The proposed solution was that contributors be asked to identify themselves as one of the following:

- 'professionals': professional botanists or biologists expert in the plants for which they are submitting observations; they have professional training or are recognized as peers by professionals;
- 'experts': do not have professional-level knowledge, but are quite experienced with using keys and descriptions, or very familiar with the plants for which they are submitting descriptions;
- 'non-expert' adults (age 18 or over) youths (12 to 18): report themselves as confident that they know the correct scientific names for the plants reported;
- child under age 12. (This last was included because it is envisioned that CalFlora may support elementary school students in sharing data among themselves.)

Contributions are not accepted from people who are even less knowledgeable than 'non-expert' as defined above.

For method of determination, contributors check all that apply from the following list:

- I recognize this taxon from prior determinations and experience;
- this determination was confirmed by herbarium specimen comparison or expert review;
- this determination included keying;
- this determination included evaluating descriptions of this and other taxa.

Finally, contributors report their certainty of an identification by choosing one of the following for each observation:[3]

- I am confident of this determination, and submit this as a positive observation of this taxon;
- I am not certain of this determination but I believe this to be a significant observation, and submit it here as an alert.

The next level of credibility assessment in CalFlora is expert review. Early additions from institutional sources were evaluated as collections, not individual records. Individual records may be reviewed and, if necessary, corrected in two ways. When the staff is alerted (via e-mail from users) to possible errors, they review individual records and make corrections as appropriate. Second, photos and occurrence records may be annotated by authorized annotators who are stringently screened for botanical skills. Subsequent viewers will see both the original record and the annotation, with the identity of the annotator.

In sum, Calflora uses four methods to address credibility: provenance, the identity and qualifications of the contributor, including his or her relationship to the taxon;

the identification method; the observer's certainty of the identification; and annotation, expert review and verification of selected observations.

All contributed observations will be stored and available. However, the observations included in default displays, those that users are likely to interpret as the 'known' distribution of a taxon, will depend on a logical evaluation of the skill of the observer, the method of determination, and the observer's certainty. Observations from 'professionals' will be included unless marked as uncertain. Other observers' observations will be excluded from default displays unless they are confirmed by an expert annotation, or include a suitable combination of determination methods and are marked as certain.

Users will have access to the entire contents of CalFlora. They will be able to make their own assessments of credibility and draw their own boundaries, based on characteristics of the observation and observer. And they will have access to the CalFlora administrators' default decisions.

9.2. *Access*

The second policy area is access. Throughout its life, CalFlora has been publicly available over the Internet without charge. CalFlora's draft policy statement on access cites the American Library Association's Code of Ethics and its policies of freedom of access to information. 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 limit access to developers, effectively excluding individuals and non-profit organizations seeking to protect the environment.

However, some members of the CalFlora governing group have expressed fear that making photos and field location information readily available over the Internet may result in over-collection or destruction of populations of rare taxa. A lengthy discussion ensued about registering users to block, or at least to discourage, people who might misuse the data. Both practical and philosophical barriers were cited. The proposed solution has been to fuzz 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 requests for suppression of location information for specific taxa.

10. *Analysis*

Networked information facilitates the crossing of social and technical boundaries in ways that Egan and Shera could not have envisioned when they called for more attention to macroscopic bibliographic methods. Knorr Cetina argues that epistemic cultures are constitutive of our knowledge society, but that there has not yet been enough attention to the inner workings of their varied epistemic machineries. She does not address the converse of this, the need for more attention to the co-ordination of epistemic machineries, the movement of information across epistemic cultures. I argue that networking of digital information has made these machineries more visible, in part by instantiating them in system design, and in part by facilitating the flow of information, including previously closely-held data, among epistemic cultures. It has at least cracked opened the black box on some of the practical solutions to the problems of shared work and credibility.

It is the purpose of this paper to make at least slightly more visible the epistemic machinery of a field heavily dependent on data sharing. Our goal is to facilitate for the design of systems of graphic communication, specifically digital libraries and other networked information systems, by means of a situational analysis of a specific instance of biodiversity data sharing.

Epistemic trust is necessary for knowledge work. Trust, we discovered, works both ways: both the user and the provider of biodiversity information potentially distrust one another when the future of the natural environment is at stake and economic and political interests are in play.

A major source of trust is shared membership in a knowledge community, whether we take a strong or weak social view of the roles of community and individual as knowers (Poutanen, 2001). Epistemic communities help participants decide who and what is credible. They have mechanisms of assessing and demonstrating competence and honesty, and, more generally, shared understandings and orientations. On the simplest level, members of the same epistemic community share methods of work, tools and language; they tend to perform similar tasks in similar ways, with common standards. CalFlora reports the methods by which taxon identifications were made as a key criterion for assessing credibility. According to Lave and Wenger and Knorr Cetina, communities of practice and epistemic cultures shape the understandings and identities of participants. Knowledge is located in the nexus of participants, practices, artefacts and social arrangements. A key element in assessing credibility, then, is determining or demonstrating membership in a community of practice. CalFlora and other arenas of 'citizen science' demonstrate that this is not simply a matter of training and formal credentials. The concerns about misuse of data show that credibility is not simply a matter of competence and lack of intention to deceive; it includes 'virtue', shared orientation and values and appropriate purposes and activities.

Scientific and other kinds of knowledge work also require the crossing of boundaries. I have argued elsewhere (Van House, 2002) that digital libraries like CalFlora are boundary objects, both in their use by varied communities and their creation by coalitions of users, information owners, and technologists.

This empirical case provides insight into how these participants both draw boundaries and cross them. CalFlora's participants are brought together across disciplines and levels of expertise by a shared commitment to conservation of plant diversity. CalFlora crosses boundaries by accepting contributions from non-expert as well as expert sources. It addresses the credibility of data from varied sources to be used for varied purposes by providing information about the observer, the observation (method), and the relationship between the observer and the observation (the observer's expertise with that taxon and certainty about that identification), and by adding a layer of expert annotation over non-expert observations. It both allows users to draw their own boundaries and makes default decisions about which observations to use.

The other kind of inclusiveness, inclusiveness in access to data, has posed more intractable problems. Preventing misuse of data was much easier when information was less available and shared through interpersonal connections. CalFlora has found it difficult prevent misuse by classifying users; the alternative is limiting overall access to sensitive data, which excludes legitimate users as well. Some have suggested that CalFlora limit use to a small group of registered users, but the preponderance of CalFlora's decision-makers believe that the benefits of public understanding of plant diversity outweigh the potential harm.

On-going innovation in graphic communication and bibliography, in networked information and digital libraries, poses new problems of credibility and testimony, and makes more visible the negotiations by which knowledge groups grapple with these problems. More situational analyses, more ethnographically-informed research into epistemic cultures, will help us understand knowledge processes generally and digital library design and use in particular.

Acknowledgement

I would like to thank the CalFlora staff, Ann Dennis and Tony Morosco, and the CalFlora Advisory Committee for their openness and warm cooperation on this project. This research was supported in part by the UC Berkeley Committee on Research through its Research Bridging Grant programme.

Notes

1. This description is as of January, 2002. Design changes have been made.
2. CalFlora has a synonymous table that shows nomenclature correspondences. Even this is not a perfect solution, since the boundaries between taxa sometimes change.
3. The screen shows this note:

> If possible, seek confirmation of determinations before submitting observations to CalFlora. Submit uncertain determinations only if you believe the observation to be highly significant and time sensitive. Method and certainty of determination, and skill group in whihc you registered, affect how your data will be displayed in CalFlora, but all information you submit will be available to the public.

References

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

Bowker, G. C., 2000, Biodiversity datadiversity. *Social Studies of Science*, 30, 643–683.

Burbules, N. C., 2001, Paradoxes of the web: the ethical dimensions of credibility. *Library Trends*, 49, 441–453.

Callon, M., Law, J. and Rip, A., 1986, How to study the force of science. In M. Callon, J. Law and A. Rip (eds) *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World* (London: Macmillan Press).

Cohen, I. J., 2000, Theories of action and praxis. In B. S. Turner (ed.) *The Blackwell Companion to Social Theory* (Malden, MA: Blackwell Publishers), pp. 73–111.

Davenport, E. and Cronin, B., 2000, The citation network as a prototype for representing trust in virtual environments. In B. Cronin and H. B. Atkins (eds) *The Web of Knowledge: a Festschrift in Honor of Eugene Garfield* (Medford, NJ: Information Today Inc. & The American Society for Information Science), pp. 517–534.

Donath, J., 2000, Being real: questions of tele-identity. In K. Goldberg (ed.) *The Robot in the Garden: Telerobotics and Telepistemology in the Age of the Internet* (Cambridge, MA: MIT Press), pp. 296–311.

Egan, M. E. and Shera, J. H., 1952, Foundations of a theory of bibliography. *Library Quarterly*, 22, 125–137.

Friedman, B., Kahn, P. H., Jr and Howe, D. C., 2000, Trust online. *CACM*, 43, 34–40.

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

Fujimura, J., 1992, Crafting science: standardized packages, boundary objects, and 'translation'. In A. Pickering (ed.) *Science as Practice and Culture* (Chicago: University of Chicago Press), pp. 168–211.

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

Gambetta, D., 1988a, *Trust: Making and Breaking Cooperative Relationships* (New York: Basil Blackwell).

Gambetta, D., 1988b, Can we trust trust? In D. Gambetta (ed.) *Trust: Making and Breaking Cooperative Relationships* (New York: Basil Blackwell), pp. 213–237.

Giddens, A., 1984, *The Constitution of Society: Outlined of the Theory Structuration* (Berkeley, CA: University of California Press).

Giddens, A., 1990, *Consequences of Modernity* (Stanford, CA: Stanford University Press).

GIERYN, T. F., 1999, *Cultural Boundaries of Science: Credibility on the Line* (Chicago: University of Chicago Press).
HARAWAY, D. J., 1991, *Simians, Cyborgs, and Women: the Reinvention of Nature* (New York: Routledge).
JONES, K., 2000, *Trust. Routledge Encyclopedia of Philosophy*. Available online: http://www.rep.outledge.com/philosophy/articles/entry/L/L107/L107.html
KNORR CETINA, K., 1999, *Epistemic Cultures: How the Sciences Make Knowledge* (Cambridge, MA: Harvard University Press).
LANE, C., 1998, Introduction: theories and issues in the study of trust. In C. Lane and R. Bachman (ed.) *Trust Within and Between Organizations: Conceptual Issues and Empirical Applications* (Oxford: Oxford University Press), pp. 1–30.
LATOUR, B., 1987, *Science in Action* (Cambridge, MA: Harvard University Press).
LATOUR, B., 1999. On recalling ANT. In J. Law and J. Hassard (eds) *Actor network theory and after* (Oxford: Blackwell Publishers), pp. 15–25.
LATOUR, B. and WOOLGAR, S., 1991, *Laboratory Life: The Construction of Scientific Facts* (Princeton, NJ: Princeton University Press).
LAVE, J., 1988, *Cognition in Practice: Mind, Mathematics, and Culture in Everyday Life* (Cambridge: Cambridge University Press).
LAVE, J. and WENGER, E., 1991, *Situated Learning: Legitimate Peripheral Participation* (Cambridge: Cambridge University Press).
LAW, J., 1986, The heterogeneity of texts. In M. Callon, J. Law and A. Rip (eds) *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World* (London: Macmillan Press), pp. 67–83.
LAW, J., 1992, Notes on the theory of the actor-network: ordering, strategy, and heterogenity. *Systems Practice*, 5, 379–393.
LAW, J. and HASSARD J. (eds), 1999, *Actor network theory and after* (Oxford: Blackwell Publishers).
LUHMANN, N., 1994, *Risk: a Sociological Theory* (New York: Aldine de Gruyter).
LYNCH, C. A., 2001. 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, 12–17.
MAIER, D., LANDIS, E., CUSHING, J., FRONDORF, A., SILBERSCHATZ, A. and SCHNASE, J. L., 2001, 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).
MECHANIC, D. and MEYER, S., 2000, Concepts of trust among patients with serious illness. *Social Science and Medicine*, 657–668.
MUNRO, A., HOOK, K. and BENYON, D. (eds), 1999, *Social navigation in information space* (London: Springer).
PICKERING, A., 1992, From science as knowledge to science as practice. In A. Pickering (ed.) *Science as Practice and Culture* (Chicago: University of Chicago Press), pp. 1–26.
POUTANEN, S., 2001, How could contemporary social theory contribute to socialize epistemology? *Social Epistemology* , 15, 27–41.
PUTNAM, R. D., 2000, *Bowling Alone: the Collapse and Revival of American Community* (New York: Simon & Schuster).
RIEH, S. Y. and BELKIN, N. J., 1998, Understanding judgment of information quality and cognitive authority in the WWW. In American Society for Information Annual Meeting, Pittsburgh, Pennsylvania, 24–29 October (Medford, NJ: Information Today, Inc), pp. 279–289.
SCHIFF, L., VAN HOUSE, N.A. and BUTLER, M., 1997, Understanding complex information environments: a social analysis of watershed planning. In Digital Libraries '97: Proceedings of the ACM Digital Libraries Conference (New York: ACM Press), pp. 181–186.
SCHNEIDER, F. B. (ed.), 1999, *Trust in Cyberspace* (Washington, DC: National Academy Press).
SELIGMAN, A. B., 1997, *The Problem of Trust* (Princeton, NJ: Princeton University Press).
SHAPIN, S., 1994, *A Social History of Truth: Civility and Science in Seventeenth-Century England* (Chicago, IL: University of Chicago Press.).
SHNEIDERMAN, B., 2000, Designing trust into online experiences. *CACM*, 43, 57–59.
STAR, S. L., 1989, The structure of ill-structured solutions: boundary objects and heterogeneous distributed problem solving. In L. Gasser and M. Huhns (ed.) *Distributed Artificial Intelligence*, 2 (London: Pitman Publishing), pp. 37–54.
STAR, S. L. and GRIESMER, J. R., 1989, Institutional ecology, 'translations,' and boundary objects: amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science*, 19, 387–420.
SZTOMPKA, P., 1999, *Trust: a Sociological Theory* (Cambridge: Cambridge University Press).
VAN HOUSE, N. A., 1995, 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.
VAN HOUSE, N. A., 2002, Digital libraries and collaborative knowledge construction. In A. P. Bishop, B. Buttenfield and N. A. Van House (eds) *Digital Library Use: Social Practice in Design and Evaluation* (Cambridge, MA: MIT Press), in press.
VAN HOUSE, N. A., Bishop, A. P. and Butterfield, B., 2002, Introduction: digital libraries as sociotechnical systems. In A. P. Bishop, B. Buttenfield and N. A. Van House (eds) *Digital Library Use: Social Practice in Design and Evaluation* (Cambridge, MA: MIT Press), in press.

VAN HOUSE, N., BUTLER, M. and SCHIFF, L., 1998, Cooperative knowledge work and practices of trust: sharing environmental planning data sets. In CSCW '98: The ACM Conference On Computer Supported Cooperative Work, pp. 335–343.
VANN, K. and BOWKER, G. C., 2001. Instrumentalizing the truth of practice. *Social Epistemology*, 15, 247–262.
WENGER, E., 1998, *Communities of Practice: Learning, Meaning, and Identity* (New York: Cambridge University Press).
WILSON, P., 1983, *Second-Hand Knowledge: an Inquiry into Cognitive Authority* (Westport, CT: Greenwood Press).