

## **Beyond Ontologies: Toward Situated Representations of Scientific Knowledge**

William Pike\* and Mark Gahegan

GeoVISTA Center  
Department of Geography  
302 Walker Building  
University Park, PA 16802 USA

wpike@psu.edu\*, mng1@psu.edu

\*Corresponding Author

### **Abstract:**

In information systems that support knowledge-discovery applications such as scientific exploration, reliance on highly structured ontologies as data-organization aids can be limiting. With current computational aids to science work, the human knowledge that creates meaning out of analyses is often only recorded when work reaches publication – or worse, left unrecorded altogether – for lack of an ontological model for scientific concepts that can capture knowledge as it is created and used. We argue for an approach to representing scientific concepts computationally that reflects (1) the situated processes of science work, (2) the social construction of knowledge, and (3) the emergence and evolution of understanding over time. In this model, knowledge is the result of collaboration, negotiation, and manipulation by teams of researchers. Capturing the situations in which knowledge is created and used helps these collaborators discover areas of agreement and discord, while allowing individual inquirers to maintain different perspectives on the same information. The capture of provenance information allows historical trails of reasoning to be reconstructed, revealing the process by which knowledge is adopted, revised, and reused in a community; as a result, end users can evaluate the utility and trustworthiness of knowledge representations. We present a proof-of-concept system, called Codex, based on this situated knowledge model. Codex supports visualization of knowledge structures through concept mapping, and enables inference across those structures. The proof-of-concept is deployed in the domain of geoscience to support distributed teams of learners and researchers by encouraging greater appreciation for shared understanding.

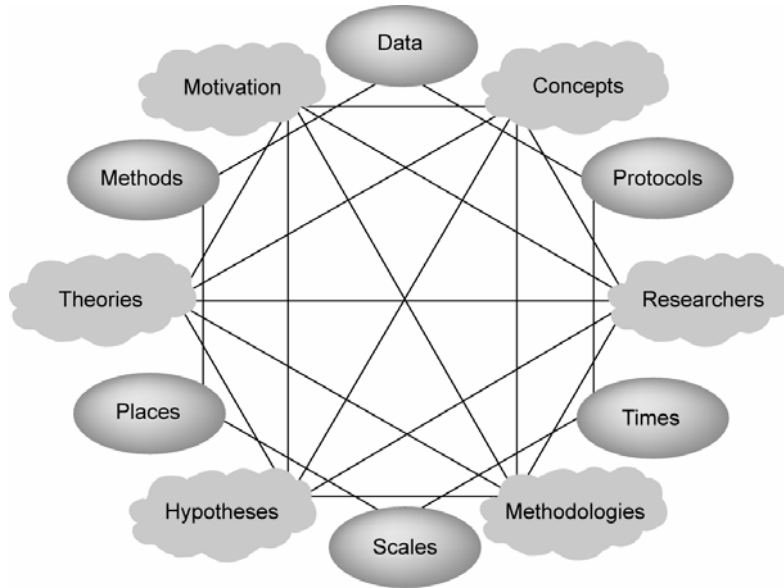
## 1. Introduction

Ontologies, as they are typically implemented in information systems, are often hierarchical and authoritative: these ontologies are useful formalizations in circumstances where formalization is called for, such as mapping terms between domains. But real-world cognition is often more fluid, flexible, and context-dependent than strict formalizations suit. Can ontologies properly capture the nuance of human knowledge? Does any single ontology reflect what is truly relevant to a particular application or domain, or is the ontology more a reflection of its creators' worldview than of neutral or common belief? Here, we propose that knowledge representations for computational environments should reflect the *situated* nature of human understanding. We also introduce a proof-of-concept application that enables distributed collaborators to share knowledge and data in a manner that is cognizant of the contexts and perspectives surrounding the creation and use of those resources.

We focus on scientific cognition as an example of human reasoning processes and on the domain of earth science as an application testbed. Like other domains, geoscience produces volumes of information, from hyperspectral sensors taking hundreds of measurements each hour, to millions of words in journal articles published every year. The scientific community's ability to generate new information – ever more detailed observations, about more diverse phenomena – often seems to outpace its ability to turn these measurements into useful knowledge. What insight was discovered and then forgotten, or discovered but never communicated? The problem is not that there is no wisdom contained in the digital artifacts of modern science, nor that contemporary science is at a standstill for its inability to make sense of increasingly complex descriptions of the world – quite the opposite, and *that* is the problem. How do we make efficient and effective use

of that knowledge? Even within a single discipline, the variety of information types and analytical methods brought to bear on a problem can complicate assessing commensurability between researchers' approaches. The clearest picture of a problem might only be painted when diverse points of view are integrated into an explanation broader than any one alone could provide. The communication of conceptual models between collaborators is crucial to accomplishing this integration, especially in earth and environmental applications (Heemskerk *et al.*, 2003).

The information science literature is rife with efforts to represent human "concepts" computationally, but the prevailing view of a concept in much research is as a category label useful for integrating heterogeneous data sources. Computational data *contains* knowledge, to be sure, and it is used to create and apply knowledge, but that knowledge is not yet represented well. As a result, information integration tasks are often data-centric; semantics are important to the extent that they support data interoperability, but the human knowledge and practices that guided the collection or use of that data remain implicit somewhere in the data's syntax or schema. A land cover map, for instance, says something about the place it depicts, although what it says to an individual researcher is either locked in the data, locked in the researcher's head, or described elsewhere in natural language text. In any case, it is not easily accessible to others who want to know how or why to use this information (say, to devise a new theory), or whether it went into any existing theories. For domains where meaning depends, in part, on the subjective perspectives of its inquirers, a restricted view of what constitutes a concept does not do justice to the complexity of human knowledge structures. Figure 1 depicts the elements of



**Figure 1.** Nexus of constructs concerning the development and application of scientific knowledge. Some (ellipses) are often made explicit in scientific reports or metadata, while others (clouds) are not; the latter, however, are crucial to understanding, communicating, and reusing scientific knowledge.

meaning as an interconnected nexus of attributes, only one of which is the concept proper, but all of which are required to fully appreciate the nature of human knowledge.

This work approaches the problem of capturing, storing, and communicating scientific knowledge by treating science foremost as a process. Knowledge is constructed and applied during this process as observations are collected and manipulated, hypotheses generated and tested, and results transmitted and built upon. Here, concepts rather than datasets are the primitive elements of scientific inquiry. This approach emphasizes interoperability of ideas, not simply data; it recognizes that the knowledge these ideas embody is by turns a shared and contested conceptualization, the result of collaboration, negotiation, and manipulation by teams of researchers. Whereas modern ontology is very much concerned with Aristotelian classification (a logic of *terms*), we aim to move toward knowledge representations as logics of *inquiry* and *interpretation*. By devising a system for capturing individual perspectives on a problem, concepts can be represented as cooperatively constructed, experientially grounded, and

semantically interoperable resources capable of reflecting their evolution, in multiple contexts, over time. Ultimately, the scientific record can be made more useful to collaborators across space and over time, as the audit trail that is captured can result in more robust explanations and lessen the likelihood of repeating dead ends.

## **2. Merging Top-down with Bottom-up**

There are two broad approaches to the problem of knowledge representation. The ontological approach is characterized by projects such as Cyc (Guha and Lenat, 1991), a sort of top-down, authoritative encyclopedia. Ontological tools such as this focus mainly on enabling sharable underlying representations of knowledge and less on interfaces and supporting infrastructure to let collaborators construct this knowledge together. The alternative approach emphasizes the bottom-up, discursive nature of knowledge. This approach acknowledges the perspectives of collaborating inquirers (rather than an imposed ontology) in defining concepts relevant to a community. The cooperative approach is evident in computer-mediated communication methods such as the Delphi method (Turoff and Hiltz, 1996), where the aim is to generate shared understanding (or areas of disagreement) over time. Cooperative tools focus on effective interfaces to collaborative work, but at the expense of creating underlying representations of that work that are not interoperable and cannot be repurposed in other systems.

### *2.1 Asserting Knowledge from the Top-down*

When a strictly ontological approach to knowledge sharing is taken (e.g., Bozsak *et al.*, 2002; Sugumaran and Storey, 2002), the end user is often left with what come across as neutral concept structures isolated from the collaborative experiences that informed their creation and from the fluid process of their change. Indeed, there have been attempts at “national knowledge

infrastructures” built from static ontologies and designed to be deployed wholesale across diverse domains and applications (Cao *et al.*, 2002). Some ontological frameworks have been extended to represent the process of constructing natural science knowledge through experimental procedures (Noy and Hafner, 2000), although these ontologies are intended to retrieve process descriptions from existing text.

There have also been some commendable recent efforts to examine the problem of ontology versioning and change as it relates to maintaining logical consistency (notably Klein *et al.*, 2002), but there is room for a deeper exploration of how changes resulting from the manipulation of resources over time can be reflected in KR. A change in the way we express a concept (through its intension, extension, or relations) over time reflects something deeper than the straightforward relabeling; the change indicates a shift in ourselves that necessitates modification to the interpretive stance others must take to understand our knowledge (Buzaglo, 2002).

It is possible to suggest that domains could agree on semantics by committee, but the success of this approach is clearly limited by diversity of opinion. Kazic (2000) suggests a middle ground, where domains create ontologies for only the most abstract, simple concepts. The simplicity of this approach is also its shortcoming: those ideas “most likely to engender controversy are left where they belong – as the private opinions of people, databases, or algorithms” (Kazic, 2000). But it is these controversial ideas, opinions, hypotheses, and theories that are often important to forming, evaluating, and modifying scientific explanations. Ideally, our ability to represent semantics computationally should not be reduced to the lowest common denominator upon which we can all agree. Disagreement may identify topics ripe for breakthrough.

There is now growing recognition in the knowledge representation field that its tools should reflect the situated work practices of their users (Schultze and Boland, 2000) and accommodate the dialogical, interactive nature of exploration (Nake and Grabowski, 2001; Dustdar, 2004). Marcos and Marcos (2001) argue that ontologies in information science are often treated as unassailable schema for “external” knowledge rather than as representations of shared knowledge with their own context and schema. Magnani (2001) suggests that *situatedness* is precisely what makes abduction a useful model for computer-based hypothesis creation – even under conditions of hypothesis failure, it produces useful information. And in light of the deductive nature of ontology languages like OWL, some have suggested a turn toward a “Pragmatic Web” instead that explicitly enables communities to test, refine, and implement emergent, rather than top-down, solutions (de Moor *et al.*, 2002).

## *2.2 Building Knowledge from the Bottom-up*

The bottom-up, cooperative approach to knowledge construction is characterized by the tools and methods of Computer-Supported Cooperative Work (CSCW). CSCW applications for scientific collaboration often take the form of electronic notebooks, organized into hierarchies of chapters and pages (e.g., Lysakowski and Doyle, 1998; Myers *et al.*, 2001), in which researchers can enter and search for free-form records (although these notebook are still linear in structure). The descriptive nature of CSCW relaxes many of the normative constraints of formal knowledge representations.

Recently, the CSCW community has begun to embrace ontologies as the basis for tools to support scholarly discourse. These ontologies describe the kind of entities that a CSCW system is capable of expressing (van Bruggen *et al.*, 2003) and are typically considered generic containers for scientific work (e.g., Suthers, 1999), not as choices of perspective that are themselves contestable. Some CSCW tools, like ScholOnto (Buckingham Shum *et al.*, 2000), develop discourse ontologies to express connections between researchers and published topics. However, the public record of science tells only part of the story, and not always faithfully; it does not reveal all of the analysis procedures, decisions, wrong turns, and intermediate results that underlie the work that merits publication. Publications are a high-level mechanism for knowledge transfer within a large community, but within science teams actively working on problems together, publications are not the primary means of communication (although they may provide background information). Much of the discourse relevant to science is thus inaccessible outside of the small groups in which it occurs. Practitioners in other places or times can have difficulty in reconstructing the discursive process that lead to a particular finding.

### *2.3 The Best of Both*

Between the ontological and the cooperative views of knowledge sharing lies an opportunity for combining machine-readable, standardized representations of knowledge with the ability for communities to elicit and refine them over time. While ontological representations of knowledge (at various levels of generality, from task to domain (Guarino, 1997)) have been described as no less than a “silver bullet” (Fensel, 2001) for information integration – and have been rapidly accepted in the information science community – there has been relatively little reflection on how conceptual structures emerge from practice and how they can reflect the evolving nature of that practice. The top-down imposition of an ontologist’s domain model masks understanding

how practitioners themselves construct meaning in ill-structured scientific problems where there may not even be initial agreement about nature of the domain itself. Usually, it is the experts in ontology who determine how to represent a domain, not communities of practitioners. Elusive concepts in geoscience, such as hazard risk, are not easily described using static ontologies, let alone ontologies that are intended to apply across an entire domain. Methods from the cooperative work community, on the other hand, address the bottom-up nature of knowledge construction but generally lack semantic richness; the expressions of meaning they produce are not routinely grounded in knowledge representations that allow concepts to be efficiently shared, searched, and reused in other problems or by other tools. The present study brings cooperative construction and emergence to ontologies, and richer semantics to cooperative tools.

### **3. The Construction of Knowledge**

Knowledge is the information that results from the accumulation of experience and reasoning, by human, machine, or both. More than just awareness of information, it involves aspects of understanding, or the ability to apply information – consciously or otherwise – to solve a problem. The distinction between awareness and understanding is a critical one, but in computational environments, awareness is sometimes conflated with understanding. A database of facts, for example, is sometimes called a “knowledge base.” But does this accumulation of facts reflect understanding (that is, are the experiences and reasoning facilities capable of making use of this knowledge present)? Or are the facts meant solely to facilitate the recollection or creation of knowledge by their user? Only in the former case could this computational information, as it is stored, properly be called knowledge. Central to this study’s approach is an

effort to incorporate aspects of understanding into the representational medium itself. Thus, the representation explicitly preserves the sense of utility that creates knowledge out of information.

### *3.1 The Ingredients of Knowledge*

We define three components that are required to represent knowledge in a more contextualized fashion: concepts, metadata, and situations. The *concept* expresses the existence of an abstract category and encompasses everything in its extension. A given concept may have different names in different circumstances while preserving the same underlying meaning (its intension).

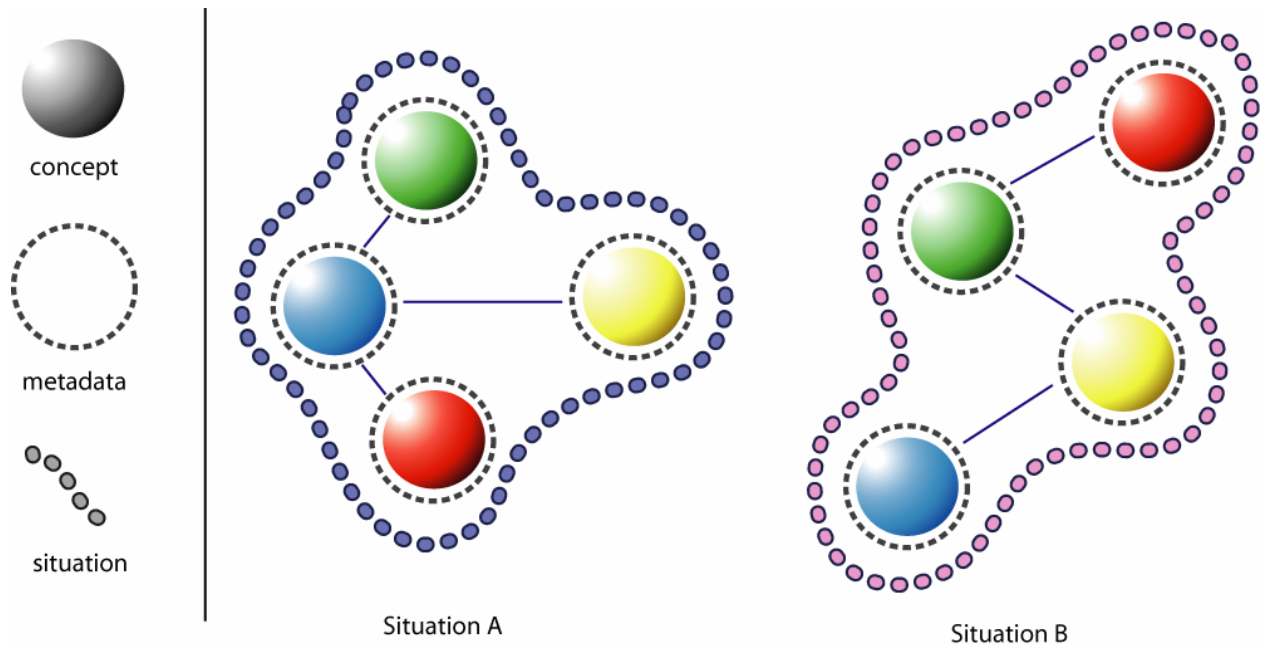
There are a number of basic views on the form of mentally held concepts. Classical Aristotelian theory holds that concept membership is defined through a set of singly necessary and jointly sufficient conditions, although it is widely recognized that this view deals poorly with “typicality effects” (the case that one conceptual member is a better example of the class than another) and that consistent definitions are scarce in practice and limit the capacity for conceptual change (Wittgenstein, 1953). In dealing with these shortcomings, the probabilistic approach to conceptual structure suggests that some concepts are better examples of their category than others. This approach finds support in empirical classification studies (e.g., Rosch, 1975; Rosch and Mervis, 1975) that indicate classification is not done on the basis of defining features so much as through proximity to prototypes. Finally, a “conceptual atomist” approach defines concepts through relationships between mental symbols and the objects they represent, a position in which concepts themselves are devoid of intensional definitions (Fodor, 1998).

A key problem for the present work is to implement a model for concepts that can be communicated efficiently across the human-computer interface. We could choose to use natural language terms, and indeed most studies of the structure of concepts focus on lexical concepts, but to effect rich knowledge representations it is desirable to move beyond simple syntactic labels for concepts. Although concepts may be difficult to define internally (that is, their intension may be vague), it is possible to describe them in terms of their relationships with other concepts. To this end, the present work bases its representation of concepts in a dimensional variety of probabilistic model. In this approach concepts are defined through the values (or range of values, as in Gardenfors (2000)) they occupy along continuous dimensions (Smith and Medin, 1981). Each *dimension* represents another concept, with an indicated *value* describing the nature of the relationship. A further characteristic of the view of concepts taken here follows from the notion of perceptual-functional affordances (Tversky, 2005) initially developed to account for visual and spatial properties of an entity. If, however, we use a dimensional approach to represent concepts, then these concepts come to occupy a multidimensional “concept space” within which we might look for some of the same functional affordances (the roles a concept plays or the capabilities it enables). Representing concepts’ functional roles in a larger knowledge structure is important to depicting “how” and “why” in scientific reasoning.

In our model, each concept is wrapped in metadata that consists of the attributes that can be recorded regardless of how or why a concept is used: who created it, using what tools, at what time and place, and so on. In the process of inquiry, concepts are selected based on relevant criteria and linked together into larger structures. These acts of conceptual manipulation have been described as *situation* (Solomon *et al.*, 1999), the bringing together of background

information and current observations and analyses toward some goal. Situation is important to knowledge representation because it explicitly reproduces the enactment that is part of selecting and reasoning with a set of concepts (Barsalou, 2002). Lemke (1997) calls situation an “ecology,” a term that evokes the dynamic interaction between concepts and thinkers in the process of knowledge construction.

Situation, then, encompasses the coordinated activity that is directed toward some goal. A given concept – we might think of it as a node in a conceptual network – can be reused in different circumstances, but there will be some information we want it to carry with it regardless of circumstance (this we have called metadata), and some that will be unique to the role it plays in a particular case (this we have called situation). To denote the particular choice of concepts, metadata, and situations that a particular thinker (or community of thinkers) uses to describe a process, problem, or phenomenon, we can use the term *perspective*. In Figure 2, situations A and B might represent different perspectives on a problem taken by two thinkers. Each might use concepts from the same body of shared understanding but see them as being directed toward different explanations. For thinker A to appreciate B’s perspective, it is necessary to reproduce for A both the entities that are relevant (the concepts and contexts) as well as their surrounding situation (the directed aim of B’s reasoning). This work creates an infrastructure for achieving that reproduction and a visual mechanism for depicting concepts and situations.



**Figure 2.** Metadata in our model describe the circumstances surrounding the creation or use of an individual resource or concept (a node in a conceptual network); situations describe the circumstances of larger knowledge structures arising from the different ways these nodes can be connected.

#### 4. Implementing a Situated Knowledge Model

In this section we describe the basic architecture for a system to support collaborative and situated knowledge representations. The convergence of the notions of situation, community, and scientific process can motivate more complete models for knowledge-centered computing than are currently available. In support of such philosophically and cognitively informed solutions, John Sowa writes that *“independently developed, but convergent theories that stand the test of time are a more reliable basis for standards than the consensus of a committee”* (Sowa, 2000). To represent our thought structures, we should not start by adopting a language like OWL as the Received Pronunciation for our domain, asking only “How can we represent our ideas using this structure?” Instead, we should ask, “What are the fundamental characteristics of knowledge-based work, and how should the design of computational tools follow from them?”

The intent of this turnabout is to show that the solutions developed below are not dependent on any one technology, but on a longstanding tradition of inquiry into human understanding.

Scientific notebooks are a familiar touchstone for thinking about recording measurements, results, and hypotheses as they occur. Leonardo da Vinci's notebooks might be considered archetypes of the genre. In the pages of these notebooks are observations of the natural world, tentative theories, diagrams and explanations of experiments. The notes are not a linear narrative; a single notebook contains insight into dozens of domains, connected by interwoven themes. Each page is bordered by marginalia that provide commentary, revisions, and links to other areas of thought. Today, these notes provide crucial insight into theories of the world from half a millennium ago. We call this style of manuscript a *codex*, a book that consists of sheaves of wood or parchment.

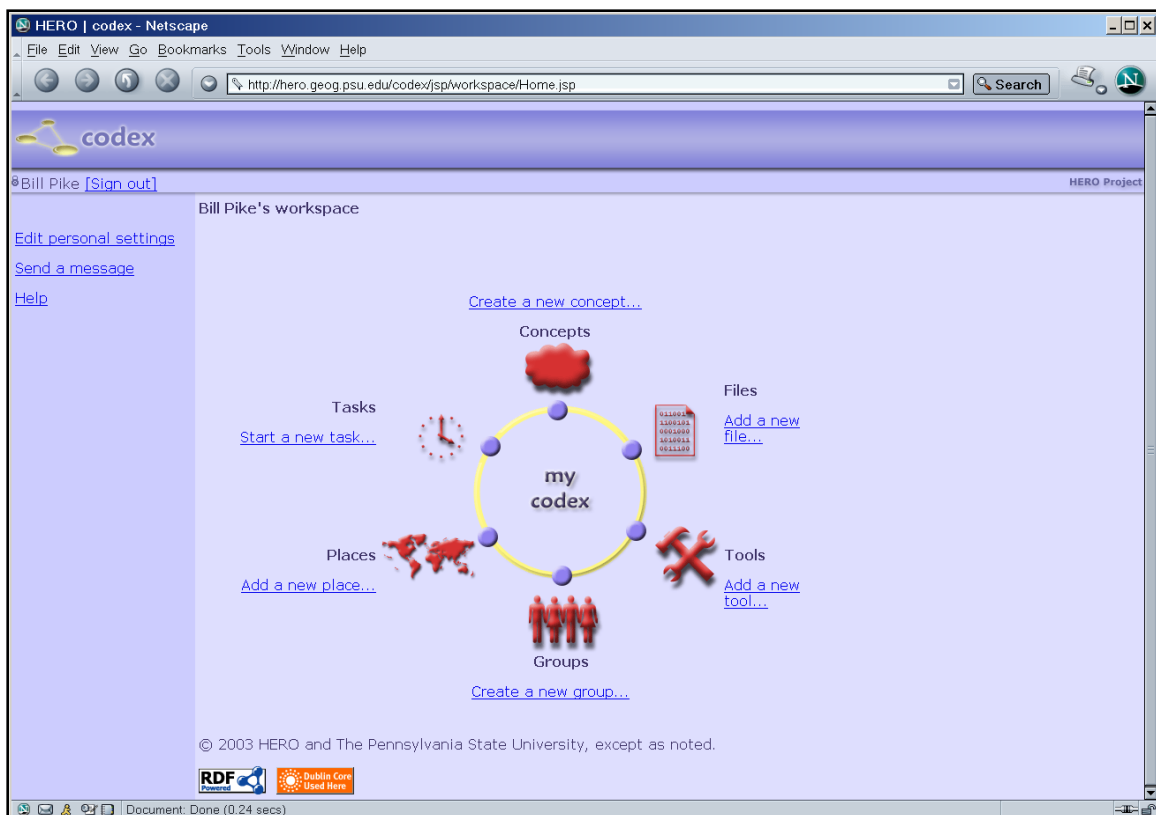
We have developed a Web-based application called Codex as proof-of-concept platform for capturing and sharing situated knowledge representations. Codex uses a portal model to organize distributed resources under a single interface. (Here, the term *resource* is a generic label for a unit of information contained in Codex; a resource might be a binary data file, a description of an abstract concept, or even a representation of a collaborator.) The portal makes it appear to the user as though the resources displayed coexist together, but in the construction of a particular view into the portal, information may have been pulled from many locations.

Codex is principally a knowledge-sharing medium, and it occupies a niche not considered by other information systems. It operates at a level of abstraction different from that of CSCW tools

intended to support data-sharing; while Codex allows data files to be stored and linked together, data are described foremost by the human concepts they signify. Codex also builds on online scientific workbenches (e.g., Stevens *et al.*, 2003) that emphasize data integration for automated analysis; Codex treats problem-solving as an issue of human consideration and interpretation. A scientific problem-solving environment might integrate several analysis tools in an attempt to support hypothesis generation (Sanchez and Langley, 2003), but fails to leverage the history of resource manipulations that result. Codex is at once a CSCW tool that enables rich semantic descriptions, and a semantic markup platform that relaxes the constraints of common ontological approaches. Codex is built around the concept of workspace. Workspaces can be both private and communal. Each Codex user has a personal workspace to store his or her ideas, data, hypotheses, and so on. Researchers can move resources to shared workspaces where they can be accessed, applied, or modified by collaborators. The cooperative, Web-based nature of Codex means that one user's insight can be made immediately accessible to colleagues a world away.

The researcher logging in to Codex is first presented with a nexus-like view onto the workspace (Figure 3). This starting point groups resources together under a set of default categories, providing quick access to the basic units of an investigation. From the workspace home page the researcher can rapidly upload a file, look in on collaborators, or describe a new analysis. Six types of resource are supported on this page, although these entry points can be supplanted with user-defined categories.

- **People.** The individuals and groups who create or apply resources accessed through the Portal. Each person maintains a profile that can communicate elements of his or her background and expertise.
- **Concepts.** Descriptions of abstract ideas, such as “flood” or “earthquake”.
- **Files.** Binary data that express something about a concept. Files could include spreadsheets, text documents, images, audio clips, maps, or other data formats (quantitative or qualitative) that connect observations or measurements to the cognitive structures represented by concepts.
- **Tools.** The methods used to analyze data and to construct instantiations of concepts (categories) from data. Tools could include GIS operations, visualization methods,



**Figure 3.** The home page for a Codex user's workspace.

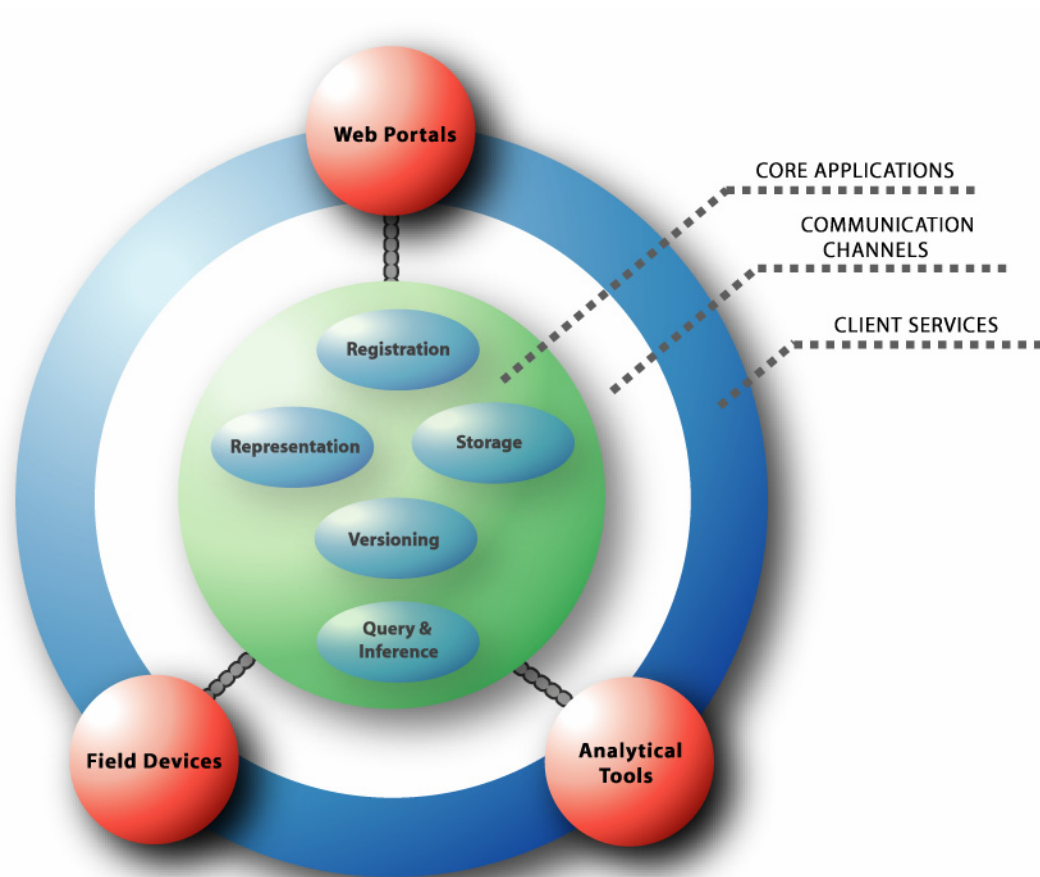
predictive models, interviewing instruments, or statistical tests.

- **Places.** Geography is fundamental to integrative research, and places help researchers define the locations and scales under study, whether described as bounding polygons or as place names. Place also helps to account for differences in epistemology between researchers.
- **Tasks.** People, concepts, files, tools, and places are linked together through tasks that might describe a workflow process, an experimental procedure, or a problem-solving approach

A nexus view into these categories underscores the preeminence of interrelations between resources types in Codex; knowledge structures express the interdependence of data, analyses, hypotheses, and results. For instance, Codex might reveal a thread showing how a *file* was produced by a particular *person* as a step in a *task* aimed at describing the relevance of a *concept* to a *place*. Codex emphasizes that the act of making connections is central to knowledge construction.

Codex is built on a two layer client-server design (Figure 4). At Codex's core is a set of server-side applications that manage the basic functionality for maintaining a shared knowledge base. Client applications, such as Web portals, mobile devices, and analytical models, interact with this knowledge base through HTTP and XML communication standards. By enforcing a separation between core functionality and client interfaces, the set of services that manages situated, perspective-based representations of scientific knowledge can be used by multiple applications at once. These applications can all be thin clients, leaving the server to do the

heavy lifting of storage and inference while the clients provide interfaces or extensions to the core functions. This architecture also lets users interact with the same knowledge base through a variety of interfaces, supporting the insight that can come from examining information in different formats. Third party clients can also be developed over time that layer domain-specific views over the same underlying applications (for instance, a Codex customized for geoscientists). Currently, the Web portal has received primary development attention, although some work on enabling access to Codex functionality through mobile devices has also been conducted.



**Figure 4.** Codex architecture.

There are five components to the Codex server.

- **Representation.** All resources in Codex are described in the OWL Full semantic markup language. The representation module maps between resource expressions in Codex clients and their corresponding OWL primitives, using the Jena API.<sup>1</sup> This module is the hook for all user interfaces to Codex; it provides a set of Java methods to interact with OWL resources (e.g., add a new resource, express a new relationship, display an existing resource) while hiding the implementational details of OWL. Consequently, neither users nor client developers need to be facile with OWL, and in fact OWL could be replaced by a next-generation knowledge representation language with no impact on a client's interface or operation. We use OWL primarily because it offers interoperability with a wide community and can be extended to suit the needs of a situated knowledge model.
- **Storage.** Concept and workspace files are stored as OWL text files to make them accessible to Semantic Web crawlers that index knowledge resources. Since the knowledge structures built with Codex are intended to be shared and reused, it makes sense to ease third party search and retrieval. Representing workspaces in OWL (that is, not only the *content* of a workspace, but the definition of the Codex workspace itself) means that researchers can use any third-party, OWL-compliant tool to access and manipulate their personal resources. The storage module is also responsible for maintaining links to external resources (for instance, OWL files imported into a workspace but that actually reside on a remote server).
- **Registration.** Each resource belongs to one or more workspaces, which represent sandboxes in which individuals and groups work. The registration module tracks the

---

<sup>1</sup> <http://jena.sourceforge.net>

assignment of resources to these workspaces and the promotion of resources from one workspace to another (e.g., from a private space to a shared space as a resource gains approval by a community). The registration component also validates user permissions, determining what resources should be exposed to a user or a search engine (keeping hidden, for instance, those tentative knowledge structures that investigators do not yet wish to share).

- **Query and Inference.** Storing resources is one thing; finding them is another. The query module uses Jena's inference engine to return resources based on their semantic relationship to the search criteria. For instance, a user interested in a particular concept can query for tasks in which it was used. The inference engine finds instances of tasks that contain the concept, perhaps limiting the results to only cases where the concept played a certain role specified by the user.
- **Versioning.** Codex resources change through use, so it is not possible to keep just one copy of a resource for all collaborators to share. Each user might make slight modifications that conform to his or her perspective. For each modification, Codex spawns a new version of the resource that contains a reference to its immediate predecessor (or predecessors, if it was created by merging properties from several resources). By following these ancestral paths, audit trails emerge that show the steps taken to put a resource into its current form. The versioning module also provides the facility to undo changes (by reverting to a previous copy of the resource) while preserving the evidence that they were made. These tentative paths that were later abandoned can still be informative.

The modular architecture of Codex leaves open the possibility of integration with other components of a knowledge representation infrastructure. For example, Codex could serve as the knowledge management node on a broader network, handling the capture and communication of explanations that emerge from manipulating information in a larger online workbench.

## **5. Modeling Knowledge in Codex**

The Concept (capital C) is the universal set in Codex; every resource and set of resources that can be described using Codex is either a member of the class of Concepts or a member of a proper subset. The six resource categories introduced above represent five proper subsets, or specializations, of Concept: file, group, place, task, and tool. The sixth, concepts, contains direct members of the class Concept (such as “chair”). The reason for this top-level category is that it allows certain rules to be instituted for the format of conceptual information and simplifies reification of resource collections as instances of another resource type (for instance, a file, place, and tool can be gathered and reified as a task). Each of the specializations extends the default Concept with unique characteristics; the file subset, for instance, adds properties for file location, size, type, and so on; the place subset accommodates attributes like place name and geographic coordinates; membership in the group subset is limited to instances of people.

The use of Concept as a universal quantifier also places Codex’s knowledge model in explicit opposition to contemporary style. Conventional wisdom holds that the ontology is the top-level category, the container for all knowledge represented computationally. This viewpoint is in fact hard-coded into the OWL language. Each OWL file is intended to represent an ontology – a

single way of structuring resources. This “ontology-first” model is highly restrictive; it presupposes a structure where there may be none (indeed, the act of manipulating resources in a tool like Codex might be for the very purpose of devising this structure, so imposing one at the start would be futile). Given the emphasis of the present work on structuring knowledge from the bottom up through pragmatic use, Codex takes a “Concept-first” approach.

Each Concept in Codex is an OWL statement that contains intensional, extensional, and contextual components. In Codex, intension is based on a dimensional variety of probabilistic model, compatible with the models of Rosch and Mervis (1975) and Gardenfors (2000): a Concept  $C$  is the set of properties  $\{P_1 \dots P_n\}$  that characterize it. Each  $P$  is another Concept typecast as an OWL property. A concept’s extensional set may be empty, or it may contain one or more instances that represent cases of  $C$  typecast as an OWL individual. A resource’s versioning history is included in its contextual markup.

Codex maintains one-to-many relationships between Concepts and situations. A situation is an arbitrary group of resources and the relationships that connect them; a given resource can be used in any number of different situations. Codex does not explicitly represent situations through a special set of tags. Representation of situational meaning is instead software-driven, based on the semantic and contextual relationships already stored within resources. The capacity to work with multiply-situated resource descriptions makes Codex unique among CSCW and ontological approaches to knowledge management.

The justification for not storing situations as explicit collections in Codex is twofold. First, there can be redundancy between situations (one resource stored many times, once for each situation in which it exists), and avoiding the storage of extraneous information is desirable. Second, situations can in theory be inferred around any set of resources. In the long run, it is wiser to store the procedures for creating collections through an arbitrary set of parameters (therefore, as software rules in Codex) than it would be to hard-code each possible collection separately.

Codex supports two varieties of situation, *user-defined* and *inferential*. A user-defined situation is formed on the basis of resource selection and/or definition by an individual researcher. For instance, in the course of defining an “Earthquake risk” concept, the researcher might:

1. Define two new concepts, “Earthquake risk” and “Distance decay.”
2. Find an existing concept, “Geographic area” and create a new instance of it, “Fault zone.”
3. Relate “Earthquake risk” to “Fault zone” through a distance decay property.

There is now a situation that contains a small set of concepts and relations. Should another user query for “Fault zone,” Codex can show that in one situation, a fault zone is a geographic area prone to earthquake risk.

Inferential situations result from detecting relationships between the contextual elements of the resources in a given set. What makes them special is that they do not require resources to have any predefined semantics; relationships between resources are inferred on the basis of co-occurrent context attributes (Langley *et al.*, 2002). Codex allows users to search for inferential situations over sets bounded by (1) the resources contained in a given workspace or (2) the

resultset of any query over a larger set. Inferential situations can be useful for spurring hypothesis generation by presenting candidate knowledge structures to the user. These structures are not ones that he or she created, but that represent other ways in which the resources in one's workspace could be connected.

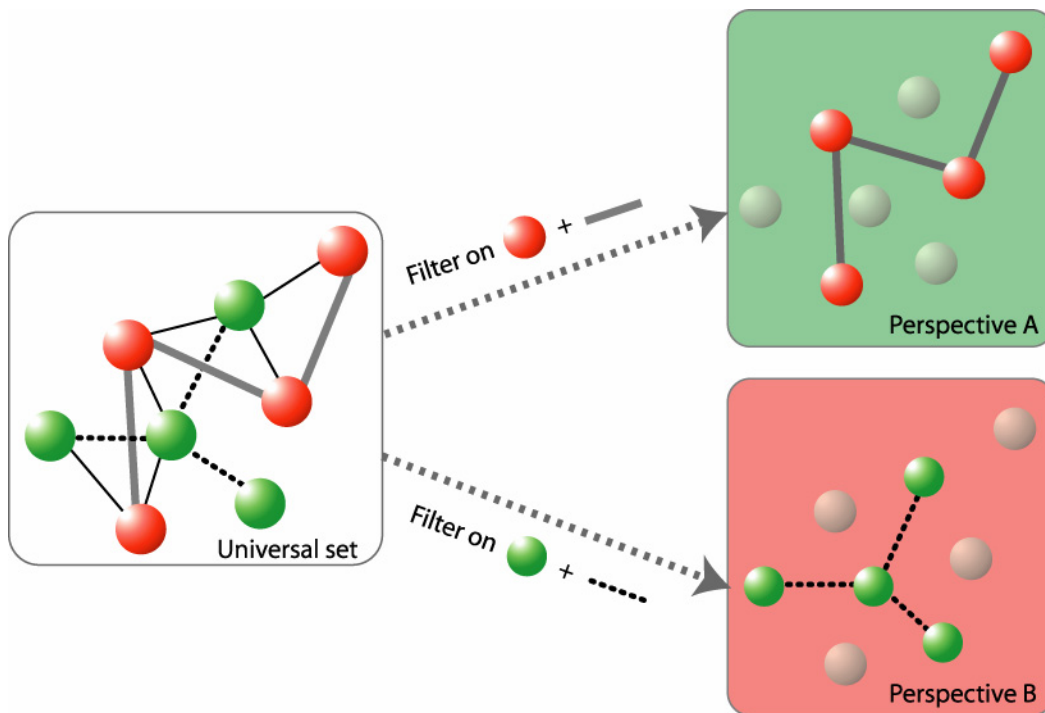
Exposing inferential situations in Codex is a variation on collaborative filtering, a self-organizing approach to detecting *ad hoc* relationships (Ansari *et al.*, 2000). Collaborative filtering is typically used to produce recommendations in e-commerce environments – for instance, based on the books for which a customer has expressed a preference, an online bookseller will recommend other books by examining what others who share those preferences have read. The same connections are possible in Codex, where an inferential situation can be built around any subset of resources that share a time, place, creator, and so on. The collaborative nature of Codex means that one can look across a user community to find relevant resources. In the simple case, Codex could build a situation around a user's query for resources that his or her collaborators created within one day of the time a target Concept was created, creating a situation of temporal association, or could recommend one researcher's "Seismology" resource to someone who already uses that person's "Earthquake risk" concept.

To understand how situations are presented to the Codex user as *perspectives* on an information space, consider the children's amusement where a colored plastic lens is passed over a complex background; suddenly a pattern appears, often a word or image in answer to a riddle. The lens absorbs certain wavelengths of light while permitting others to pass through. The result is that some of the complexity of the printed background is obscured, allowing only the salient elements

– those that are compatible with the composition of the lens – to be seen. A Codex perspective works on the same model (Figure 5). The perspective filters out some information, revealing only certain “wavelengths” of meaning that conform to the resource types present in a given situation. The remainder of a concept space is masked.

To examine a set of resources from different perspectives, the Codex user foviates on a resource and queries for the situations in which that resource is found. Codex can combine situations to either restrict or expand the selection of resources salient to a perspective.

- In the **union** of situations, the researcher finds the bounds of a problem space, given by the complete set of resources that a community deems relevant to it.
- The uniqueness of a particular perspective is found in the relative **complement** (or set



**Figure 5.** Perspectives filter a complex information space according to particular situations. Perspectives A and B preferentially select different types of resources and relations from the universal set of all Codex resources.

theoretic difference) between situations. That is, the uniqueness of a researcher's perspective can be described as the set of resources that are in the situation through which he or she describes a problem, but that are not found in anyone else's. Taking the complement of a perspective can also reveal areas of disagreement or uncertainty, where concepts in one user's perspective fail to correspond to those in others'.

- The resources and relations in an **intersection**, the consilient set, constitute a new situation that represents the points of agreement within a community. Codex uses the consilient set as the basis for expressions of community or domain belief that might qualify to be used elsewhere as top-down knowledge structures (e.g., ontologies).

Although perspectives can be compared and integrated in Codex, Codex does not mandate the use of a "neutral" ontology as might be the case with other tools. A neutral ontology amounts to a mapping vocabulary that regulates interoperability between terms. In Codex, a common vocabulary can be *discovered* if one exists, but it is not an *a priori* requirement to describe knowledge in Codex.

### 5.1 *Creating and Using Situations through Concept Mapping*

Codex's primary user interface is a concept mapping utility developed around an open-source dynamic graph browser.<sup>2</sup> Figure 6 shows a sample Codex concept map developed around a set of information resources related to the notion of seismology (the examples in this section and the next are drawn from users in the GEON project,<sup>3</sup> a large cyberinfrastructure project for the geosciences for which Codex is being developed). In the client, OWL classes or individuals are

---

<sup>2</sup> TouchGraph: <http://www.touchgraph.com/>

<sup>3</sup> <http://www.geongrid.org>

depicted as nodes; relationships between nodes (i.e., their properties) are depicted as edges. The semiotic nature of concept representations in Codex is seen here in the form of iconic representations for selected nodes. For instance, the concept “Seismic reflection” (which this user has chosen to signify through a particular data set) stands for a particular reflection profile (the *object*), which is pictured. The graph structure itself is one situation in which the resources it contains are found.

As a user draws a graph, he or she is creating a situation for its set of resources. When the user adds a copy of an existing resource (whether as a node or an edge), Codex now contains multiple situations for that resource: the situation in which it was originally created, and the new situation in which it is being applied. Each of these situations constitutes a different perspective on the same resource (either the perspectives of different researchers, if the resource has been used by

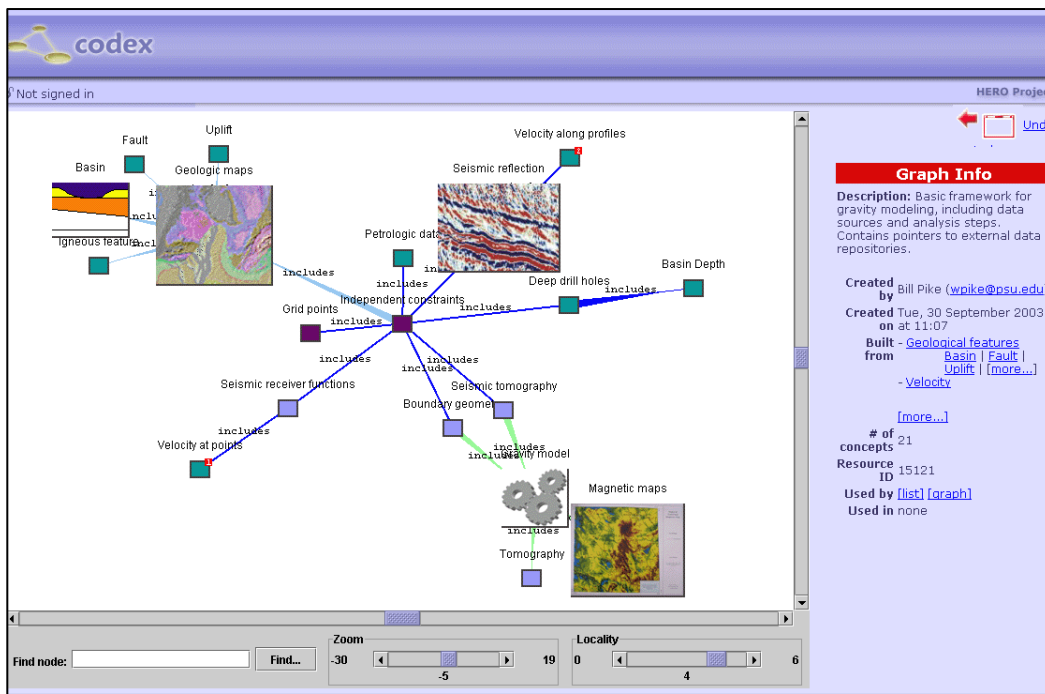
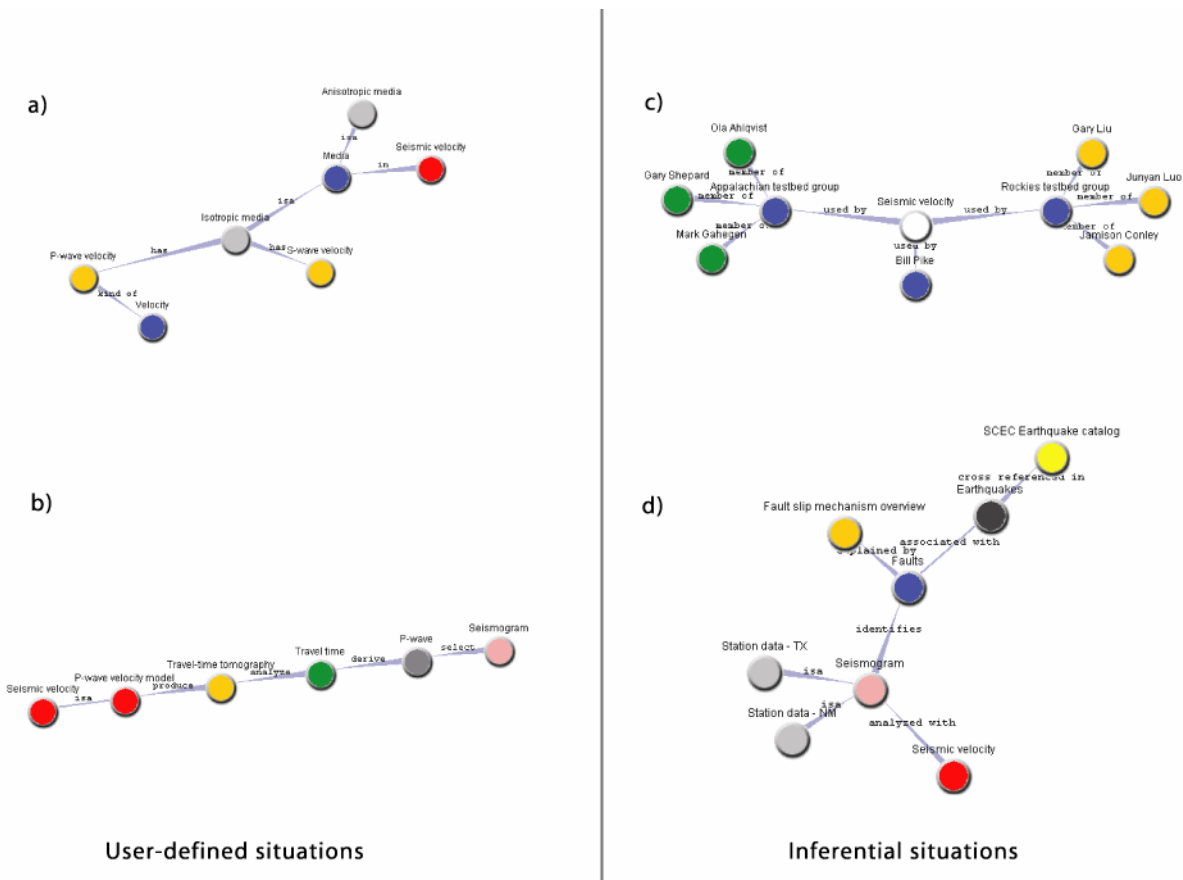


Figure 6. Codex concept map client.

multiple people, or the various perspectives that an individual scientist adopts when thinking about the resource in different situations). In facilitating reuse of existing resources, Codex enables users to extend their understanding by borrowing ideas from collaborators. This mechanism is a variation on the network elaboration technique (Eckert, 1998), which has been shown to be especially useful in pedagogical settings: those learning a new domain can start with a simple structure provided by an instructor, text, or colleague and gradually extend it with new information as their learning progresses.

A simple example shows how different perspectives are displayed in the Codex concept map client. Suppose a geoscientist has created a concept map describing the domain of seismology



**Figure 7.** Four perspectives on a “seismic velocity” concept (red node). a) Intensional concept structure. b) A task that describes how seismic velocity can be measured. c) A social network built around users of the concept. d) Data resources that have been used to describe seismic velocity.

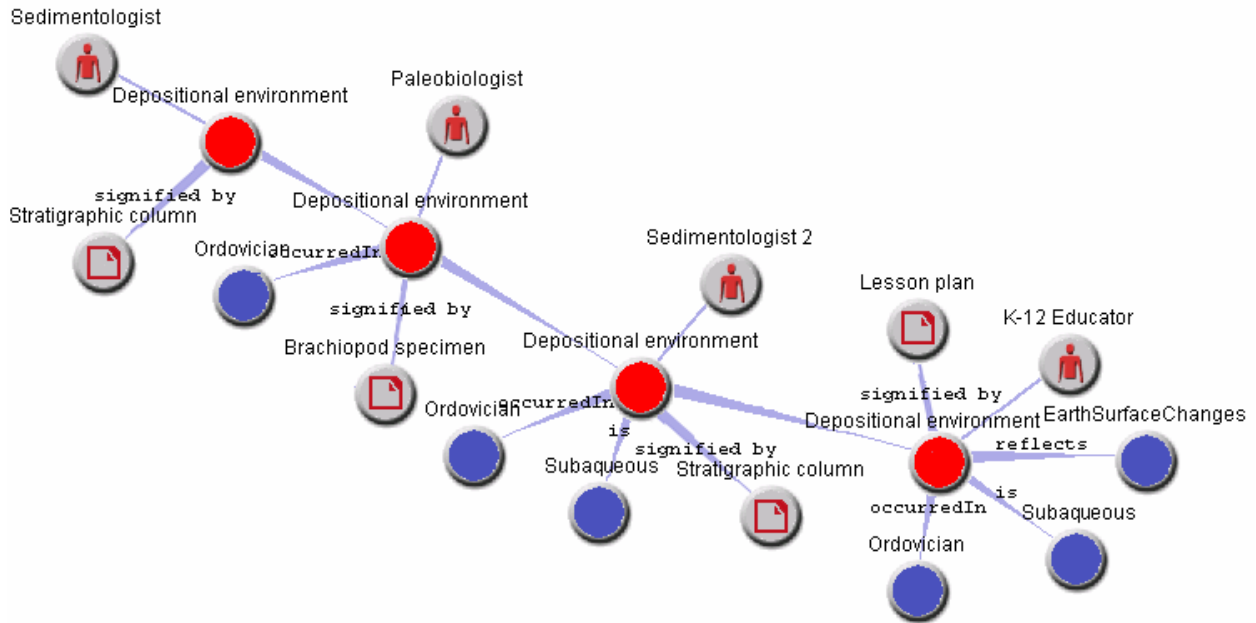
(Figure 7a). In this example, the graph represents a taxonomy and contains only concepts, so we would say that this situation is directed toward describing the *structure* of a domain. Now that the user has described this structure, he or she is interested in finding other situations for one of the concepts it contains, “seismic velocity.” (It is possible to find situations for a set of resources of any size, but for simplicity we will use a single node). The user selects the seismic velocity node and queries for a “task” perspective. Codex searches for situations in which a task has been described that includes this concept; Figure 7b reveals such a task. Now the user has gained a new perspective on the concept of seismic velocity, seeing it situated in a “how-to” network. Next, the user might want to know which collaborators have also used this seismic velocity concept; executing this query produces Figure 7c; based on information in context headers, Codex finds users who have applied the concept. In this case, two of the “users” are in fact groups that have applied the concept in shared workspaces, so the researchers shown use the resource indirectly through their membership in the groups. Finally, the geoscientist wants to know what data files contain information on seismic velocity. Running this query produces the perspective on seismology shown in Figure 7d; here, nodes represent instances of files that other researchers have included in the extension of seismic velocity. While the first two situations are user-defined (they only exist if a Codex user has built them), the latter two are inferential – they can be built automatically.

By enabling users to quickly shift perspectives on a problem, the visual display of information in Codex changes to correspond with a user’s cognitive focus. Rather than show all possible relations between seismic velocity and any other resource, the perspective filters facilitate interpretation by displaying only the nodes that are salient to the chosen situation. Changing the

visualization to suit the user's goals has been shown to increase the efficiency and effectiveness of interpretation (Neuwirth *et al.*, 1998); in light of to the fluid, adaptive nature of problem-solving, enabling this sort of change can quickly communicate multiple approaches on a problem to a thinker coming to grips with its complexity (Chung *et al.*, 2003). The Codex concept map client and its mechanism for navigating across perspectives is a step toward increasing the transfer of understanding among collaborators.

### *5.2 Combining Perspectives and Versioning to Track Knowledge Evolution*

When versioning histories and registration of concepts with different perspectives are combined, it becomes possible to present a single view onto the evolution of a resource. Figure 8 shows such a view as retrieved by Codex, summarizing the intensional and extensional changes in the geoscience “Depositional environment” concept (red nodes). The original concept, at upper left, was created by one sedimentologist and contained a single extensional element. As successive researchers adopted the concept, the connections they made between this concept and those in their personal ontologies were added to the concept's intension (blue nodes). For extension (gray nodes with document icon), practitioners are permitted to use different data to signify the same concept, or the same data to signify different concepts; there is none of the one-size-fits-all registration between data and ontology that is common in other systems.



**Figure 8.** Evolution of “Depositional environment” concept through use by researchers in different communities of practice, progressing from upper left to lower right.

## 6. Conclusions

We argue for the use of knowledge representation techniques capable of reflecting the situated nature of human cognition. By representing knowledge in the contexts of its creation and use, it is possible for these situated representations to integrate both ontological (top-down) and discursive (bottom-up) approaches to knowledge elicitation and structure. By capturing knowledge as it is constructed evolves, this work draws from the benefits of computer-supported cooperative work. By allowing groups to treat certain concept structures as representative of their community (and to share them as such), this work acknowledges the utility of top-down ontologies and the structure they can lend to a problem.

The benefits yielded by situated knowledge models and their implementation in collaborative systems should be felt across the range of inquiry-based activities. At the outset, pedagogy and

basic scientific research are obvious application contexts – students and researchers are actively engaged in (and indeed have as part of their remit) knowledge exposition and communication. In these applications, a tool like Codex formalizes, extends, and speeds an extant knowledge transfer process. Still, there are possibilities for a Codex-like approach in other application domains. Indeed, in any field where information is reused, built upon, or acted upon, the representation of the situated reasoning that went into a resource can help guide appropriate use. Planning and policymaking tasks, for instance, can require searching for consilience in group perspectives over time. In strategy-setting endeavors, understanding the coherence of individual perspectives and detecting common threads can lead to policy candidates likely to have widespread approval. (Although Codex, like policy, need not be democratic; particularly distinctive perspectives on a problem can be detected, appreciated, and adopted or rejected by others). In competitive intelligence, systems that help analysts create meaning out of disparate information resources can support effective sense-making. A knowledge model that preserves audit trails of resource manipulation and concept growth can increase the transparency of a research enterprise; audiences can engage in deeper critical examinations than prima facie reports might ordinarily allow. Furthermore, the trustworthiness of any information resource can derive from its coherence with a “reference” perspective – thus, trust need not be a consistent measure for all circumstances, but can vary according to the needs of an inquirer or the conditions of a situation.

The professional culture of many scientific domains, however, creates significant barriers to adopting a knowledge sharing system such as that proposed here. When accolades, promotions, and other measures of status are often attained through ruthlessly protecting one’s own

intellectual property, a sea change in science would be required to create a culture of open sharing. This work asks scientists to confront such a change, and this change could have clear benefits – not least a greater ability to integrate observations and hypotheses across space and over time – but without institutional support (from academic departments, funding agencies, research laboratories, journal publishers, and so on) its success could be hard won. But there are early signs that the need to achieve community synthesis is changing the way science is performed; the US National Science Foundation, for instance, is beginning to encourage greater knowledge sharing through community ownership of resources produced by its funding. NSF requires that software products under its National Middleware Initiative use open-source licenses. But the techniques suggested in this paper will have benefit even if only adopted at the individual, not collaborative, level. Indeed, it is possible that the benefits individual researchers might realize from using these tools (combined with the weight of a funding agency like NSF) will encourage the cultural shifts necessary to make their collaborative use possible.

## **Acknowledgements**

The material presented in this paper is based upon work supported by National Science Foundation grants EAR/ITR-0225673, BCS-9978052, and BCS/ITR-0219025.

## **References**

- Ansari, A., S. Essegiaier and R. Kohli (2000). Internet recommendation systems. *Journal of Marketing Research* **37**: 363-375.
- Barsalou, L. (2002). Being there conceptually: Simulating categories in preparation for situated action. *Representation, Memory, and Development: Essays in Honor of Jean Mandler*. N. Stein, P. Bauer and M. Rabinowitz (eds.). Mahwah, NJ, L. Erlbaum: 1-16.
- Bozsak, E., M. Ehrig, S. Handschuh, A. Hotho, A. Maedche, B. Motik, D. Oberle, C. Schmitz, S. Staab, L. Stojanovic, N. Stojanovic, R. Studer, G. Stumme, Y. Sure, J. Tane, R. Volz and

- V. Zacharias (2002). KAON - Towards a large scale Semantic Web. *E-Commerce and Web Technologies, Proceedings*. Lecture Notes in Computer Science. **2455**: 304-313.
- Buckingham Shum, S., E. Motta and J. Domingue (2000). ScholOnto: An ontology-based digital library server for research documents and discourse. *International Journal on Digital Libraries* **3**(3): 237-248.
- Buzaglo, M. (2002). *The Logic of Concept Expansion*. New York, Cambridge University Press. 182 p.
- Cao, C. G., Q. Z. Feng, Y. Gao, F. Gu, J. X. Si, Y. F. Sui, W. Tian, H. T. Wang, L. L. Wang, Q. T. Zeng, C. X. Zhang, Y. F. Zheng and X. B. Zhou (2002). Progress in the development of national knowledge infrastructure. *Journal of Computer Science and Technology* **17**(5): 523-534.
- Chung, P. W. H., L. Cheung, J. Stader, P. Jarvis, J. Moore and A. Macintosh (2003). Knowledge-based process management - an approach to handling adaptive workflow. *Knowledge-Based Systems* **16**(3): 149-160.
- de Moor, A., M. Keeler and G. Richmond (2002). Towards a pragmatic web. *International Conference on Computational Science, Proceedings*. U. Priss, D. Corbett and G. Angelova (eds.), Springer-Verlag. Lecture Notes in Computer Science. **2393**: 235-249.
- Dustdar, S. (2004). Caramba - A process-aware collaboration system supporting ad hoc and collaborative processes in virtual teams. *Distributed and Parallel Databases* **15**(1): 45-66.
- Eckert, A. (1998). The "Network Elaboration Technique": A computer-assisted instrument for knowledge assessment. *Diagnostica* **44**(4): 220-224.
- Fensel, D. (2001). *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. New York, Springer. 138 p.
- Fodor, J. (1998). *Concepts: Where cognitive science went wrong*. New York, Oxford University Press. 174 p.
- Gardenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA, MIT Press. 307 p.
- Guarino, N. (1997). Understanding, building, and using ontologies. *International Journal of Human-Computer Studies* **46**: 293-310.
- Guha, R. V. and D. B. Lenat (1991). Cyc - a Midterm Report. *Applied Artificial Intelligence* **5**(1): 45-86.
- Heemskerck, M., K. Wilson and M. Pavao-Zuckerman (2003). Conceptual models as tools for communication across disciplines. *Conservation Ecology* **7**(3).
- Kazic, T. (2000). Semiotics: A semantics for sharing. *Bioinformatics* **16**(12): 1129-1144.
- Klein, M., D. Fensel, A. Kiryakov and D. Ogyanov (2002). Ontology versioning and change detection on the Web. *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*. A. Gomez-Perez and V. Richard Benjamins (eds.). Berlin, Springer-Verlag. Lecture Notes in Computer Science. **2473**: 197-212.
- Langley, P., J. Shragar and K. Saito (2002). Computational discovery of communicable scientific knowledge. *Logical and Computational Aspects of Model-Based Reasoning*. L. Magnani, N. Nersessian and C. Pizzi (eds.). Amsterdam, Kluwer.
- Lemke, J. (1997). Cognition, context, and learning: A social semiotic perspective. *Situated Cognition: Social, Semiotic, and Psychological Perspectives*. D. Kirshner and J. Whitson (eds.). Mahwah, NJ, Erlbaum: 37-55.

- Lysakowski, R. and L. Doyle (1998). Electronic lab notebooks: Paving the way of the future of R&D. *Records Management Quarterly*: 23-28.
- Magnani, L. (2001). *Abduction, Reason, and Science: Processes of Discovery and Explanation*. New York, Kluwer. 205 p.
- Marcos, E. and A. Marcos (2001). A philosophical approach to the concept of data model: Is a data model, in fact, a model? *Information Systems Frontiers* **3**(2): 267-274.
- Myers, J., E. Mendoza and B. Hoopes (2001). *A collaborative electronic notebook*. Proceedings of the IASTED International Conference on Internet and Multimedia Systems and Applications, Honolulu.
- Nake, F. and S. Grabowski (2001). Human-computer interaction views as pseudo-communication. *Knowledge-Based Systems* **14**: 441-447.
- Neuwirth, C. M., J. H. Morris, S. H. Regli, R. Chandhok and G. C. Wenger (1998). *Envisioning communication: Task-tailorable representations of communication in asynchronous work*. Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work, Seattle, ACM Press. 265-274.
- Noy, N. F. and C. D. Hafner (2000). Ontological foundations for experimental science knowledge bases. *Applied Artificial Intelligence* **14**(6): 565-618.
- Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology* **104**(3): 192-233.
- Rosch, E. and C. Mervis (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* **7**: 573-605.
- Sanchez, J. and P. Langley (2003). *An interactive environment for scientific modeling and discovery*. Proceedings of the International Conference on Knowledge Capture, Sanibel Island, FL, ACM Press. 138-145.
- Schultze, U. and R. J. Boland (2000). Knowledge management technology and the reproduction of knowledge work practices. *Journal of Strategic Information Systems* **9**: 193-212.
- Smith, E. and D. Medin (1981). *Categories and concepts*. Cambridge, MA, Harvard University Press. 203 p.
- Solomon, K., D. Medin and E. Lynch (1999). Concepts do more than categories. *Cognitive Science* **3**(3): 99-104.
- Sowa, J. (2000). Ontology, metadata, and semiotics. *Conceptual structures: Logical, linguistic, and computational issues*. B. Ganter and G. Mineau (eds.). Berlin, Springer-Verlag. Lecture Notes in Computer Science. **1867**: 55-81.
- Stevens, R., A. Robinson and C. Goble (2003). myGrid: Personalised bioinformatics on the information grid. *Bioinformatics* **19**(Suppl. 1): i302-i304.
- Sugumaran, V. and V. C. Storey (2002). Ontologies for conceptual modeling: their creation, use, and management. *Data & Knowledge Engineering* **42**(3): 251-271.
- Suthers, D. (1999). *Representational support for collaborative inquiry*. Proceedings of the 32nd Hawaii International Conference on System Sciences, Maui, HI, IEEE. 1076.
- Turoff, M. and S. Hiltz (1996). Computer based Delphi processes. *Gazing into the Oracle: The Delphi Method and its Application to Social Policy and Public Health*. M. Adler and E. Ziglio (eds.). London, Kingsley.
- Tversky, B. (2005). Form and function. *Functional Features in Language and Space*. L. Carlson and E. van der Zee (eds.). New York, Oxford University Press.

- van Bruggen, J., H. Boshuizen and P. Kirschner (2003). A cognitive framework for cooperative problem solving with argument visualization. *Visualizing Argumentation*. P. Kirschner, S. Buckingham Shum and C. Carr (eds.). London, Springer-Verlag: 25-47.
- Wittgenstein, L. (1953). *Philosophical Investigations*. G. Anscombe (trans.). New York, Macmillan. 232 p.