Supporting Collaborative Science through a Knowledge and Data Management Portal

William Pike, Ola Ahlqvist, Mark Gahegan, Sachin Oswal

GeoVISTA Center, Department of Geography
Pennsylvania State University
University Park, PA 16802
{wpike, oka1, mng1, sachinoswal}@psu.edu

Abstract

This paper discusses a portal used by geoscientists and human-environment relations researchers to capture and share the evolution of concepts and the emergence of agreement through collaboration. A concept graph interface to the portal encodes relationships among people, concepts, data, tools, tasks, times, and places. By linking resources in flexible and reusable structures, we attempt to situate knowledge representation in the context of scientific practice.

Introduction

Complex scientific problems increasingly require collaboration between teams of researchers distributed across space and time. Moreover, such collaborators have different perspectives that guide their choice of methods, models, vocabularies, and philosophies. Effective collaboration remains dependent, however, on researchers' ability to co-create meaningful explanations, and most importantly on being able to describe the knowledge that supports these explanations. This paper presents an approach to leveraging the Semantic Web to enable such collaboration, allowing researchers to describe, share, and reuse both the process and the products of scientific research.

The Web tool we describe, dubbed Codex (available at http://hero.geog.psu.edu/codex) after manuscript notebooks such as da Vinci's, is aimed not just at enabling collaboration through data integration – a task that is addressed by a number of existing approaches – but also at capturing the evolution of ideas and their application, as well as their relationship to data. Our approach is to support the sharing of information resources in a way that allows their changing roles and relationships to be captured and explored as fluidly as possible.

Three Abstract Dimensions of Collaborative Scientific Computing

To achieve our central aim of facilitating knowledge sharing, coordinated activities among scientists must resolve questions concerning the distribution of resources, their agreed or contested meaning, and their location and means of access. Accordingly, tools that support the sharing of scientific resources over the Web can be categorized along the dimensions of collaboration, distribution and representation identified in Figure 1. Through collaborative work, researchers negotiate concepts and categories that cross cognitive spaces, from personal (supported by such tools as DSpace [1]) to community (e.g., ScienceDesk; sciencedesk.arc.nasa.gov). Second, through access to distributed resources, researchers use data, concepts, and tools created in other locations or for other problems. Tools such as HINTS [2] are built on centralized data and knowledge bases, while Grid-based tools such as myGrid [3] are designed around pipelines that connect dispersed data stores and analysis nodes. Most importantly, science relies on representations of phenomena and concepts through data, language, and computation; at one extreme such representation can be

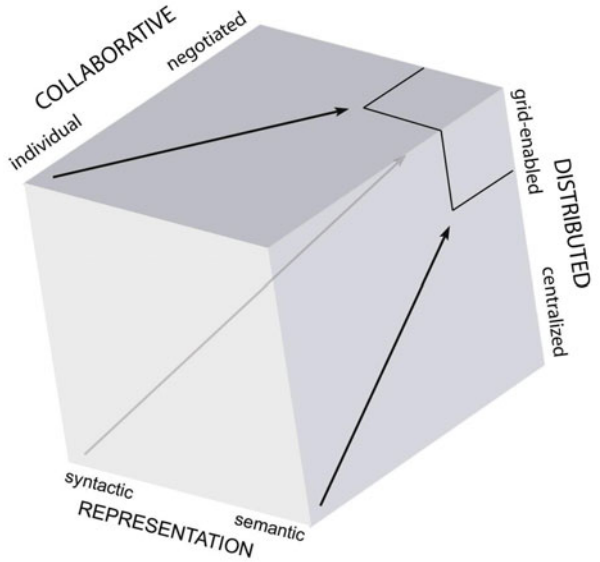

Figure 1. Computational aids to science work involve aspects of collaboration, distributed resources, and semantic representation. The upper-right corner of this space defines an area of optimal balance.

purely syntactic, as in traditional metadata standards for data products. Alternatively, applications like KAON [4] allow the construction of ontologies that afford semantic representation of concepts and their relationships.

In balancing collaboration, distributed computing, and semantic representation in Codex, we emphasize the importance of situated cognition in the practice of science. The situatedness of science work describes a context of place, time, person, problem, and perspective that informs the tactics a researcher or team takes to an analysis and greatly influences the outcomes. Even the simple act of sharing data sets can be (and ought to be) informed by this context: How were these data used? To what problems were they applied? What hypotheses or explanations were derived from these data (or what assumptions underpinned their collection)? Together, answers to these questions may help the user address how to use these data in their own problems: the semantic underpinning provides a rich basis by which to search for useful resources, to see how others have used them and to better understand their strengths and limitations. This approach to science treats it as an evolving conversation [5], in which meaning is interpreted and negotiated over time. Moreover, the nature of situations suggests that the interoperability of *ideas*, not necessarily of *data*, should be the crucial feature of any aid to science work. The Semantic Web offers us the ability to combine data markup with knowledge representation, and more specifically, *with representations of situations that help to explain and contextualize what data mean to people*.

Beyond Ontologies

Our approach to capturing the situatedness of science work centers on the notion of the concept as the primitive element in Codex. Like other tools that facilitate collaboration among scientists (e.g., [6]), we enable researchers to record the more concrete elements of research such as data sets or modeling approaches; however, Codex treats these as specializations of abstract concepts that play particular roles in scientific practice. For instance, quantitative data such as remotely sensed images do not necessarily have a meaning independent of the concepts they signify (perhaps land cover categories), while the data file itself signifies an image concept.

Ontologies as they are commonly used in information science are only one way of associating a set of concepts, and while problems can be described through ontologies at different levels of abstraction [7], the nature of an individual ontology privileges one kind of structure among a set of concepts over the fluidity with which these concepts are created, modified, and applied. As ideas emerge and evolve during scientific exploration, ontological structure may not always exist (or at least be known by a user) *a priori*. However, if concepts can be elucidated over time, their various ontological structures may begin to emerge as researchers associate them in different contexts. Initially, researchers need to be able to create structures that reflect tentative associations between concepts, and compare these evolving structures with those of collaborators to find areas of agreement. Such agreement might help communities build ontologies by revealing core concepts that have wide support, without initial resort to top-down models. We thus treat ontologies as "ontologies of convenience" that represent the coalescence of an arbitrary set of concepts around some situation – such as a task, person, place – at some point in time.

Concept representation

Recognizing the gap between a symbolic [8] approach to represent knowledge, common in ontological structures, and associationist approaches (such as were proposed by John Locke and David Hume), as often used in data mining, we base the representation of concepts in Codex on a third approach, the cognitive theory of conceptual spaces [9]. A conceptual space, intended to bridge symbolic and associationist models, is a multidimensional property space constructed from a number of properties such as temperature, shape, location, and so on. A property is defined as a point or region in a low dimensional subspace, for example the interval of lengths that is used to separate a tall person from a short person. One important aspect of this notion of a concept is that a property can itself be treated as a special case of a concept. Moreover, each property that contributes to a concept's definition is assigned a certain salience, or importance, in relation to other properties of the concept. This weight enables us to separate different perspectives on a concept by declaring certain properties more important in one context and peripheral in another.

We formally represent a concept space as a collection, or set, of property definitions. A property definition is represented as a set of values from a certain domain, for example the interval of height values. To represent the semantic uncertainty we often find in concept definitions such as "tall", we use the idea of rough fuzzy sets [10]. Work on fuzzy [11] and rough [12] extensions of traditional set theory have provided viable techniques to handle two important aspects of semantic imprecision, vagueness and indiscernibility. Fuzzy and rough set theories have since been further generalized into rough fuzzy sets, a joint representation for vague and resolution-limited information. Following Ahlqvist et al. [13], we use a pair of rough fuzzy definable sets $\left(\underline{\mu_x}, \overline{\mu_x}\right)$ to represent property values in an approximation space where an equivalence relation imposes granularity on a finite universe of discourse. Each approximation space is essentially a property that we use to define a concept; for example, length can be approximated by values 'tall' and 'short'. In this way any concept $C(S_i, R_i, W_i)$ is formalized as vectors of approximation spaces S_i with property values R_i given as rough fuzzy set definitions, and accompanying salience weights, W_i.

Rather than mandating a distinction between knowledge objects that constitute concepts and those that constitute

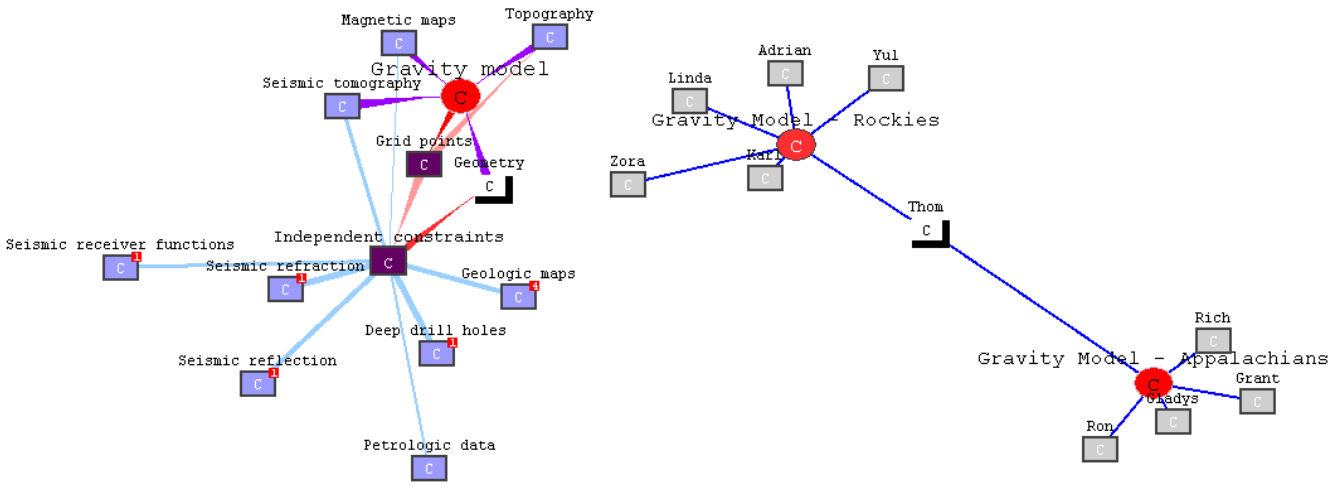

Figure 2. Two views of a gravity model concept (red nodes). An ontological description (left) shows how one geoscientist constructs such a model; a social network (right) reveals which users favor different instances of the model, with edge length suggesting the degree of support. (Concept graphing in Codex modified from open-source Touchgraph [www.touchgraph.com])

properties, we model both as concepts. As a result, a concept is defined through the interaction of other concepts that play the role of its properties. Property roles reflect relations between concepts in a given context. For example, a *mineral* concept may have *cleavage plane* as a property. A *cleavage plane* is itself a concept that might include properties such as *angle*. When considering "things that are measured with angles", a cleavage plane may be a relevant concept; the fact that it can also have the role of a property of a mineral may not be of immediate relevance.

By treating concepts as collections of properties that reflect various elements of situation (among them what problems they have been used to solve, who has used them, what data they relate to, and so on), different semantic structures can emerge by querying Codex for relationships among concepts. Such structures might include:

- **Ontologies** that relate concepts taxonomically or through task descriptions;
- **Social networks** that describe who created or used certain resources;
- **Temporal structures**, such as timelines that record events in a resource's history of use and modification;
- **Spatial structures** that reveal how resources are related in geographic or attribute space.

Figure 2 shows two sample concept structures created in Codex that demonstrate how comparison across concept properties can reveal different forms of emergent structure.

Web Implementation

Codex is designed as a Web portal (Figure 3) that provides a uniform interface to a distributed set of resources. A portal architecture frees the user from concern over where data files or concept objects are stored and how to make them interoperate. The portal model also favors personalization of a user's interaction with the system; to this end, the workspace is the organizing metaphor in Codex. A workspace contains references to all of the resources that a user has created or applied, and allows each user to customize personal views (conforming to his or her own perspectives) onto a concept space.

By default, Codex provides six entry points into a collaborative scientific environment: concepts, files, tools, people, places, and tasks. Each of these entry locations serves as a quick-access point for the full set of resources stored in all the workspaces a researcher has access to (as well as those that he or she might import from elsewhere on the Web). For example, the *tools* entry point allows users to browse the workspace from a tool-centric view, organizing information (which might include other resources such as concepts, tasks, and so on) from the point of view of the analysis or modeling tools that use these resources or have been used by them. (Currently, Codex supports a set of analysis and visualization tools from GeoVISTA Studio [www.geovista.psu.edu], but will soon support distributed Web services as well.) The default entry points are but one way of classifying different specializations of concepts; a user can configure his or her entry points by creating an index out of any property tag in any concept definition. Thus, a user could specify that one of the entry points be defined by all resources that are affiliated with a particular project or that have been used in geological modeling.

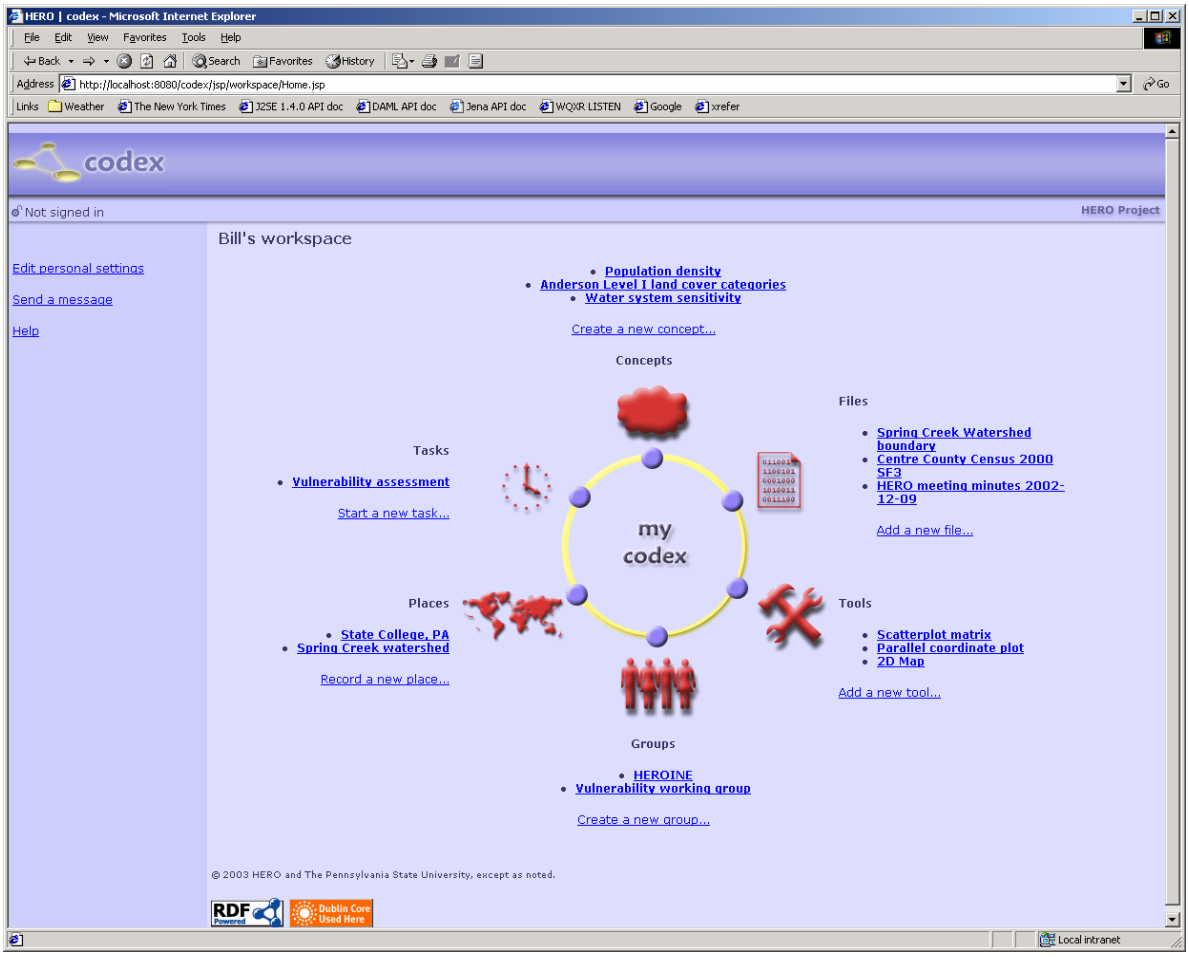

Figure 3. Portal interface to a user's personal, group, and community workspaces.

While the portal is essentially an interface to a body of linked scientific resources described in DAML, the domain users to which the portal is targeted are not ontologists per-se. As a result, special attention has been paid to the problem of knowledge capture. It is unsuitable to ask users to encode statements about resources as RDF triples, for example. Instead, users require a cognitively appropriate interface that, while exploiting DAML's semantic expressiveness, also masks the details of its implementation. While for the sake of efficient capture it may be desirable to build support for concept representation into the very tools that scientists already use, the diversity of tools even within a domain makes this impractical. Instead, we opt for the ubiquity of a Web interface, and there will also be an interface to the portal for handheld devices.

The Codex portal relies on an interactive graph visualization interface through which researchers can construct concept graphs that represent ontologies, workflow diagrams, or other networks (Figure 2 shows examples of these graphs). Such visualization approaches are commonly applied to knowledge spaces (e.g., [14]), and graph interfaces can be particularly effective for displaying relationships among different kinds of entities [15]. Here, we use graphs to display different ways of linking sets of independently defined resources. Users may map visual properties of a graph (such as shape, color, transparency, border, or label) to properties of the concept objects contained within the graph. For instance, concepts that contain a particular property might be given a certain color, or the size of a graph node might be scaled to the value of a numeric property. The use of visual properties along with dynamic navigation and expansion of graphs supports the creation of explanations and discovery of relationships [16], [17]. In addition to adding information to a workspace through concept graphs, Codex can build graphs in response to a query (an example of an ontology of convenience, the defining characteristic of which is simply that it satisfies a query).

Concepts that are shared among collaborators are continually updated as a user manipulates a graph, allowing even synchronous collaboration in the development of knowledge structures. Codex also supports interoperability with other tools by enabling any component to be serialized in DAML, including concepts, the contents of a user's workspace, and graph visualization schemes.

Supporting Collaboration

In addition to providing each user with a personal workspace, Codex facilitates collaboration over the construction and application of resources through the use of shared workspaces. Any group of users can create a common workspace for shared concepts, files, tools, and so on. Individuals can make any resource in their personal workspace available to a team, or can protect resources as private (and can even indicate whether or not they wish queries to return a resource in response to another user's question). Teams can nest, such that resources can be promoted from individual, to group, to domain workspaces as they are adopted by wider communities.

In a collaborative environment that reflects the distributed and dynamic character of science and of the researchers who conduct it, changes or differences in individual concept spaces are inevitable. Domain changes, adaptation to different tasks, or changes in conceptualization all necessitate variance. Management of these conceptual changes involves: (i) identifying an appropriate definition for a particular use of a concept and (ii) change tracking and handling the backward compatibility of revisions. Based on the work of Klein [10], Codex incorporates a framework to relate and integrate different concept definitions. This includes a versioning scheme for tracking changes to a concept and a lineage tool that logs the evolving process of knowledge creation, application, and update.

Versioning amounts to preserving temporal relations among concepts. At its most basic, versioning simply allows multiple definitions for the same concept. Different users might rely on different versions in their work, and new versions could break relationships in existing concept structures. "Breaking the build" of a concept structure is most likely when a new version removes properties that represent crucial relations in another's work. In a distributed environment, however, it can be impossible to know what concept structures might be affected by a new version, since the search space for structures that use that concept is essentially unbounded. As a result, every concept property contains references to the resources from which it was constructed, if it was not defined from scratch. Any change to a resource creates a copy of the resource to reflect the change, leaving the original intact.

Versioning affords much more than enabling different users to work with different versions of a concept without interfering with each other, however. Versioning allows users (or the system) to search over particular concept spaces (defined by a range of times, a set of users, a kind of research problem, and so on), to trace the evolution or emergence of common concepts from different resources (or, for that matter, to trace their divergence). Figure 4 illustrates different ways that new concepts can be created from existing versions. During evolution (Figure 4a), users might create new concept objects to suit their current needs by importing properties from existing concepts. For example, to define a new *mineral* object, a user might borrow properties from an existing mineral definition and add to these some new properties that modify the definition to conform to his or her perspective. In addition to specifying each property, a user can specify measurement dimensions for that property. In the case of a *temperature* property, for instance, measurement dimensions might include interval values like *Fahrenheit*, *Centigrade*, and *Kelvin*, as well as the process of mapping between them. In other cases, measurement dimensions might consist of nominal categories. For each measurement dimension, the user can specify a value or range of (potentially rough fuzzy) values for the measurement quantity itself. A researcher might assert that a particular *mineral* has a melting point *temperature* with measurement dimension *degreesCentigrade* and measurement value *2300*; alternatively, one could be less specific and assert that there is simply some mineral that has some melting point temperature, without specifying how to measure that property or what the bounds of its possible values might be.

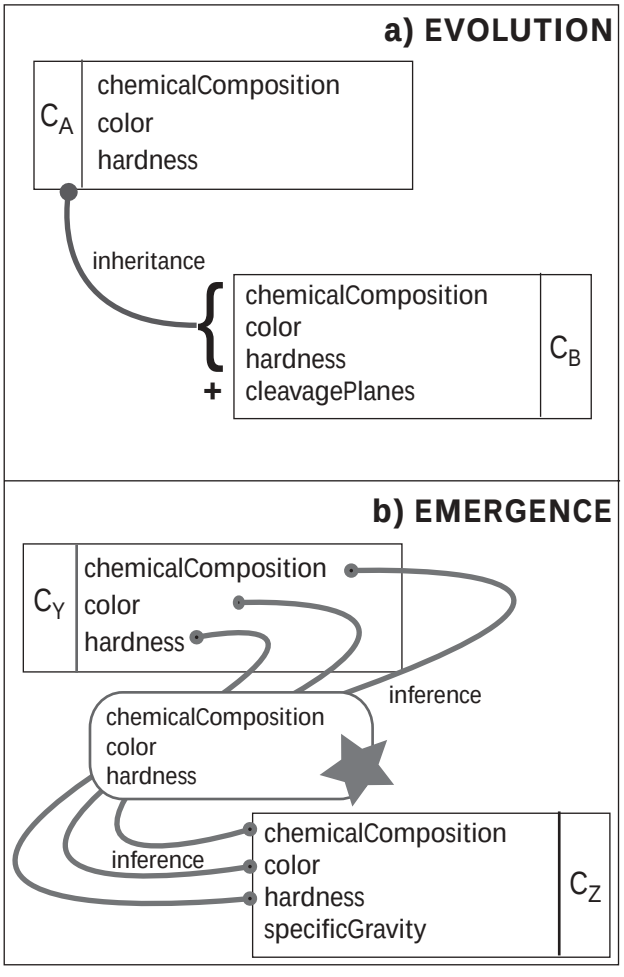

Figure 4. a) Concept C_B evolves from C_A by borrowing existing properties and adding cleavagePlanes as a new property. b) Concepts C_Y and C_Z are created independently, perhaps at different times and by different people, but a common concept (starred) emerges after they are inferred to have some degree of similarity based on shared properties.

It is also possible that concepts constructed at different times, for different purposes, or by different people, may exhibit some degree of emergent similarity. Versioning supports the detection of emergence by allowing the system to track the temporal relations among concepts in concurrent or sequential use even if they were never explicitly linked by a user. In Figure 4b, a common concept represents the overlap between different constructions. These constructions may or may not have the same label, or even describe the same things. Nonetheless, should these constructions both be used within a community, we might infer that their intersection represents some point of agreement. Of course, these properties could describe tabletops as well as minerals, so in practice concepts are further defined through properties that describe more detailed aspects of situation. A mineral might be used in the process of a rock identification task, whereas a tabletop will likely not be.

Conclusions

Currently, a team of undergraduate students from four universities around the US are using Codex to share concept definitions associated with their research on the sensitivity of local drinking water systems to environmental change (http://hero.geog.psu.edu/). The students use the portal to define individual (researcher specific), local (place specific) and community (domain specific) concepts, find points of agreement between concepts as they are constructed and applied in different locations, and link concepts to data. By creating networks of concepts based on any set of common properties, Codex is able to show these users how common conceptual structures can be constructed from independently defined ideas. The concept graphing tools are also playing a role in capturing and communicating concept maps for a number of current science projects, including the Geosciences Network (GEON: http://www.geongrid.org/).

Future challenges in this research area involve integrating analyses into concept structures such that users can directly link data to concepts through online exploration. In addition, we aim to implement a suite of semantic similarity measures to extend the tool's inferencing capabilities and more advanced concept visualization schemes to support knowledge exploration.

Acknowledgements. This research was supported by NSF grants BCS-9978052 (HERO), ITR (BCS)-0219025, and ITR (EAR)-0225673 (GEON).

References

1. Branschofsky, M. 2002. DSpace: MIT's digital repository. *Abstracts of Papers of the American Chemical Society* 224: 033-CINF.
2. Lang, K. and M. Burnett. 2000. XML, metadata, and efficient knowledge discovery. *Knowledge-Based Systems* 13: 321-331.
3. Wroe, C., et al. 2003. A suite of DAML+OIL ontologies to describe bioinformatics web services and data. *International Journal of Cooperative Information Systems* 12(2): 197-224.
4. Bozsak, E., et al., *KAON - Towards a large scale Semantic Web*, in *E-Commerce and Web Technologies, Proceedings*. 2002. p. 304-313.
5. Winograd, T. and F. Flores. 1986. *Understanding computers and cognition*. Norwood, NJ: Ablex. 207 p.
6. Myers, J., E. Mendoza, and B. Hoopes. 2001. A collaborative electronic notebook. In *Proceedings of the IASTED International Conference on Internet and Multimedia Systems and Applications*, August 13-16, 2001. Honolulu.
7. Guarino, N. 1997. Understanding, building, and using ontologies. *International Journal of Human-Computer Studies* 46: 293-310.
8. Newell, A. and H. Simon. 1976. Computer science as empirical inquiry. *Communications of the ACM* 19(3): 113-126.
9. Gardenfors, P. 2000. *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press. 307 p.
10. Dubois, D. and H. Prade, *Putting rough sets and fuzzy sets together*, in *Intelligent Decision Support: Handbook of Applications and Advances of Rough Sets Theory*, R. Slowinski, Editor. 1992, Kluwer: Boston. p. 203-232.
11. Zadeh, L. 1965. Fuzzy sets. *Information and Control* 8: 338-353.
12. Pawlak, Z. 1991. *Rough sets: Theoretical aspects of reasoning about data*. Boston: Kluwer. 229 p.
13. Ahlqvist, O., J. Keukelaar, and K. Oukbir. 2003. Rough and fuzzy geographical data integration. *International Journal of Geographical Information Science* 17(3): 223-234.
14. Chen, C. and J. Kuljis. 2003. The rising landscape: A visual exploration of superstring revolutions in physics. *Journal of the American Society for Information Science and Technology* 54(5): 435-446.
15. Kirschner, P., S. Buckingham Shum, and C. Carr. 2003. *Visualizing Argumentation*. London: Springer-Verlag. 216 p.
16. Lucieer, A. and M.-J. Kraak. 2002. Interactive visualization of a fuzzy classification of remotely sensed imagery using dynamically linked views to explore uncertainty. In *Accuracy 2002 Symposium*, 348-356. Melbourne, Australia.
17. MacEachren, A. 1995. *How Maps Work*. New York: Guilford. 513 p.