The Pennsylvania State University
The Graduate School
College of Earth and Mineral Sciences

REPRESENTING GEO-PRAGMATICS

A Thesis in
Geography
by
Boyan Brodaric

© 2005 Boyan Brodaric

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2005
The thesis of Boyan Brodaric was reviewed and approved* by the following:

Mark Gahegan  
Professor of Geography  
Thesis Advisor  
Chair of Committee

Alan M. MacEachren  
Professor of Geography

William E. Easterling  
Professor of Geography

Frederico Fonseca  
Assistant Professor of Information Sciences and Technology

Roger M. Downs  
Professor of Geography  
Head of the Department of Geography

*Signatures are on file at the Graduate School
ABSTRACT
Networks of logically defined concepts, known as ontologies, are being used as reference points for the meanings shared by digital messages in emerging cyber-based scientific information systems. This falls far short of the grander vision of such systems as a new operational paradigm for science, in which ontologies might aid knowledge discovery by serving as digital proxies for scientific knowledge. To enable scientific discovery, ontologies must additionally embrace the contextual knowledge evident in science, such as methodological histories, implicit assumptions, evolving concepts and diverse scientific perspectives.

In this research, ontologies are considered digital instruments of scientific inquiry. A contextual theory of geoscientific ontology representation is developed, formalized, and tested against scientific practice and real data. The theory takes a geo-pragmatic viewpoint, in which discovery is facilitated through development of a computable formalism for elements of the geoscientific discovery process. The knowledge represented with the formalism can be tested and replicated by geoscientists, and have sufficient context to help stimulate new hypotheses. The formalism affects the structure and content of geoscientific ontologies: (1) concept structure is augmented by including origins, effects, uses, and their production functions; and (2) ontology content is affected by the introduction of situated concepts that are defined by unique geographical histories.

The formalism is empirically tested by studying how geoscientists collaborate to develop a shared geologic map. The results are computationally implemented by using the formalism as a basis for a geospatial database schema in which the empirical data is hosted. This enables periods of geoscientific knowledge discovery and evolution to be explored visually. Geo-pragmatic influences are then used to help organize knowledge within the schema, through the introduction of several levels of abstraction for geoscience concepts including a level for situated concepts.

These studies cumulatively show that geo-pragmatic representation is realistic, viable, and valuable, as it is shown respectively to fit geoscientific data and practice, be computationally implemented, and improve current geoscience knowledge representation practices.
# TABLE OF CONTENTS

LIST OF FIGURES .............................................................................................................. VI
LIST OF TABLES ............................................................................................................... VIII
ACKNOWLEDGEMENTS ................................................................................................. IX

## CHAPTER 1: INTRODUCTION AND OVERVIEW ............................................................ 1
1.1 MOTIVATION: ENABLING INTEGRATIVE SCIENCE AND SOCIETAL IMPACT ........... 1
1.2 PROBLEM STATEMENT: GEOSPATIAL INTEROPERABILITY AND THE MEANING GAP .... 2
1.3 OBJECTIVES AND METHODS ................................................................................. 6
1.4 OUTLINE OF THE DISSERTATION .......................................................................... 8
1.5 CONTRIBUTIONS TO GISCIENCE ......................................................................... 10
1.6 EXISTING PEER-REVIEWED PUBLICATIONS ARISING FROM THIS DISSERTATION .... 11

## CHAPTER 2: CHALLENGES TO REPRESENTING GEOSCIENCE KNOWLEDGE IN CYBERINFRASTRUCTURE ................................................................. 12
2.1 GEOSCIENCE INFORMATION ............................................................................. 12
2.2 INFORMATION INTEROPERABILITY ................................................................. 13
2.3 ONTOLOGIES AND INFORMATION INTEROPERABILITY ....................................... 16
2.4 ONTOLOGIES AND GEOSCIENCE KNOWLEDGE ................................................. 22
2.5 ONTOLOGIES AND PRAGMATICS .................................................................. 28
2.6 SUMMARY OF ASSUMPTIONS AND CHALLENGES .............................................. 42

## CHAPTER 3: REPRESENTING GEO-PRAGMATICS ........................................................ 45
3.1 SCIENTIFIC PRAGMATICS ................................................................................. 45
3.2 THE NATURE OF GEO-PRAGMATICS ................................................................ 50
3.3 ELEMENTS OF GEO-PRAGMATICS .................................................................. 54
3.4 FORMALISM FOR GEO-PRAGMATICS ............................................................... 64
3.5 EXAMPLE OF GEO-PRAGMATICS .................................................................. 65
3.6 SUMMARY ........................................................................................................ 70

## CHAPTER 4: EXPERIMENTS TO ILLUSTRATE THE GEO-PRAGMATIC NATURE OF GEOSCIENCE INFORMATION ............................................................ 72
4.1 INTRODUCTION ................................................................................................. 72
4.2 BACKGROUND .................................................................................................... 77
4.3 REPRESENTATION AND VISUALIZATION METHODOLOGY .................................. 82
4.4 RESULTS .......................................................................................................... 88
4.5 DISCUSSION (I) ............................................................................................... 94
4.6 DISCUSSION (II) ............................................................................................. 97
4.7 CONCLUSIONS ............................................................................................... 98
### CHAPTER 5: STRUCTURING GIS DATABASES FOR GEO-PRAGMATICS

- **5.1 INTRODUCTION** ................................................................. 101
- **5.2 GEOSCIENTIFIC DATABASES** ............................................. 102
- **5.3 CONCEPT STRUCTURE** ..................................................... 104
- **5.4 IMPLEMENTATION** ............................................................ 108
- **5.5 OTHER IMPLEMENTATIONS BY THIRD PARTIES** ................ 114
- **5.6 CONCLUSIONS** ................................................................. 123

### CHAPTER 6: LEVELS OF ABSTRACTION IN GEO-PRAGMATIC ONTOLOGIES

- **6.1 INTRODUCTION** ................................................................. 124
- **6.2 LEVELS OF ABSTRACTION—EXISTING APPROACHES** ........ 127
- **6.3 LEVELS OF ABSTRACTION IN LEKXIS** ............................... 133
- **6.4 ORGANIZATION OF LEKXIS** ........................................... 143
- **6.5 APPLICATION OF LEKXIS** ............................................... 146
- **6.6 CONCLUSIONS** ................................................................. 149

### CHAPTER 7: CONCLUSIONS

- **7.1 SUMMARY AND CONCLUSIONS** ........................................ 151
- **7.2 CONTRIBUTIONS, LIMITATIONS AND FUTURE DIRECTIONS** ... 153

### REFERENCES.................................................................................. 161

### APPENDIX ..................................................................................... 175

- **A.1 ATTRIBUTE PREPARATION** .................................................. 176
- **A.2 SCHEMA PREPARATION** ...................................................... 181
LIST OF FIGURES

Figure 1.1: varying classification of geological regions across maps ........................................ 5
Figure 2.1: outline of chapter 2 ............................................................................................... 12
Figure 2.2: interoperability levels for databases, with examples ........................................... 14
Figure 2.3: the DOLCE upper level ontology ........................................................................ 17
Figure 2.4: integrating geoscience map databases using ontologies ...................................... 19
Figure 2.5: types of geoscientific knowledge ............................................................................. 25
Figure 2.6: a sign for the concept ‘Baker Brook basalt’ ........................................................... 29
Figure 2.7: realist semantics for ontologies ............................................................................. 34
Figure 2.8: cognitive semantics .............................................................................................. 34
Figure 2.9: cognitive-realist semantics ..................................................................................... 35
Figure 3.1: human, machine and natural pragmatics .............................................................. 47
Figure 3.2: the scientific knowledge cycle in information systems ........................................ 48
Figure 3.3: pragmatics for a geological map ............................................................................ 49
Figure 3.4: the syntactic information square .......................................................................... 51
Figure 3.5: the semantic ontological square .......................................................................... 51
Figure 3.6: the pragmatic square ............................................................................................ 51
Figure 3.7: abstraction, evolution and verification of a rock body individual ......................... 56
Figure 3.8: augmented pragmatic square with perspectives, versions and mappings .......... 58
Figure 3.9: example of the augmented pragmatic square ...................................................... 59
Figure 3.10: example of geoscientific reasoning in the pragmatic square ............................... 66
Figure 3.11: logic-based example of geoscientific reasoning in the pragmatic square .......... 67
Figure 4.1: ontologies and individual models in geological mapping .................................... 75
Figure 4.2: the study area in the Macquoid geological belt of northeastern Canada ............. 77
Figure 4.3: number of sites visited each by geologist per concept .......................................... 78
Figure 4.4: graded structure in the clusters of concepts X and Y ............................................ 81
Figure 4.5: training a SOM ..................................................................................................... 83
Figure 4.6: developing numeric values for the ‘feature’ space dimensions .............................. 85
Figure 4.7: how to interpret the intra-geologist’s thematic MMD clustering graphs .............. 86
Figure 4.8: Sammon view (right) of thematic clusters for units A, B and C ............................ 89
Figure 4.9: view of thematic spaces for concepts A and B ..................................................... 90
Figure 4.10: similarity of geologists’ thematic data in three concepts at weekly intervals .... 91
Figure 4.11: diameter of a geologist’s thematic cluster measured for concepts A, B and C . . 92
Figure 4.12: diameter of the geologists’ sampling cluster in geographical space .................. 93
Figure 4.13: distance between the prototype and geologists’ clusters for A, B, and C ........... 94
Figure 4.14: distribution of data for concept C, by geologist ................................................ 96
Figure 5.1: the traditional relations between a concept and its extension ............................... 105
Figure 5.2: conceptual schema for geoscience pragmatics, Part A ......................................... 107
Figure 5.3: conceptual schema for geoscience pragmatics, Part B .......................................... 108
Figure 5.4: physical implementation in MS-Access of the geo-pragmatic schema ............... 109
Figure 5.5: classes for: ‘Unit Cfw’, versions ‘Unit Cfw1’, ‘Unit Cfw2’, ‘Unit Cfw3’ ............... 111
Figure 5.6: classes for versions of properties of ‘Unit Cfw’ .................................................. 111
Figure 5.7: clusters for versions of properties of ‘Unit Cfw’ .................................................. 112
Figure 5.8: A context for ‘Unit C’, including an interpretation for concept ‘Unit C’ ............ 113
Figure 5.9: a blend for ‘Unit Cfw’ .......................................................................................... 114
Figure 5.10: map views of situation 1008, an instance of ‘Unit Cfw’ ..................................... 114
Figure 5.11: conceptual schema for geologic map information .............................................. 117
Figure 5.12: reclassifying map units into rock types .......................................................... 119
Figure 5.13: exploring sustainable development concepts in GBDL using ontologies........ 120
Figure 5.14: the three tier architecture of GBDL ............................................................... 121
Figure 5.15: ESRI geology data model fragment ............................................................... 121
Figure 5.16: visualizing an occurrence and concept in the experimental ESRI data model 122
Figure 5.17: logical instantiation of the map-oriented conceptual model .......................... 122
Figure 6.1: horizontal and vertical levels of ontology ......................................................... 128
Figure 6.2: three abstraction levels implied by material ontology .................................... 129
Figure 6.3: indefinite abstraction levels via logical languages ......................................... 130
Figure 6.4: three abstraction levels implied by conceptual spaces .................................. 131
Figure 6.5: a semiotic view of abstraction ...................................................................... 132
Figure 6.6: simplified example of LEKXIS in UML ......................................................... 138
Figure 6.7: extended example of LEKXIS in UML ........................................................... 140
Figure 6.8: OWL-FULL representation of part of the LEKXIS example ......................... 142
Figure 6.9: parts of a cluster and class for a situated concept ......................................... 143
Figure 6.10: semantic ISA relation and pragmatic blendMembership relations ............... 145
Figure 6.11: semantic granularity in the semantic cube .................................................... 147
Figure 7.1: applying the pragmatic square within an abstraction level .............................. 157
Figure 7.2: composition of clustering in category theory .................................................. 158
Figure A.1: the field database schema ............................................................................. 175
Figure A.2: example data for the field database schema .................................................. 176
Figure A.3: development of representative site vector $S$ from an input set $A$. .............. 182
LIST OF TABLES

Table 2.1: characterization of architectures for database networks ........................................ 16
Table 2.2: semantic and pragmatic aspects of concepts .......................................................... 30
Table 3.1: pragmatics for a scientific artefact ........................................................................ 46
Table 3.2: mappings, functions, entities and reasoning in the pragmatic square .................... 52
Table 3.3: direct and indirect of inference of objects in the pragmatic square ....................... 53
Table 4.1: recognition accuracy for supervised classification ............................................... 89
Table 4.2: supervised SOM classification on all and some thematic clusters ....................... 89
Table 4.3: framework for field-based discovery of situated concepts .................................. 97
Table 5.1: pragmatic functions and reasoning in the pragmatic square ................................ 106
Table 5.2: summary of the key constituents of geo-pragmatics ......................................... 106
Table 5.3: cluster and class for map unit property concepts .............................................. 110
Table 5.4: map unit concepts in a national database ......................................................... 118
Table 6.1: levels of ontology for substances: a comparison of approaches ......................... 134
Table 6.2: summary of LEKXIS levels ............................................................................ 137
Table 6.3: example of applying LEKXIS to the ‘Mountain’ concept ................................... 138
Table 6.4: semantic confusion matrix .............................................................................. 148
Table A.1: numerical encoding for some nominal database attribute values ...................... 177
ACKNOWLEDGEMENTS
I would like to acknowledge the agencies and people who made this work possible, and to whom I am grateful: my wife Gail, and daughters Katerina and Anne Michaela, for their constant joy and care; my parents for praying this work to completion; my advisor Mark, for his unprecedented availability, guidance, and above all tolerance; my committee members for not losing hope; my employer, the Geological Survey of Canada, for the opportunity and financial assistance in the first place; the US Geological Survey for its support; and finally my many professional colleagues who showed me the way.
CHAPTER 1: INTRODUCTION AND OVERVIEW

1.1 Motivation: enabling integrative science and societal impact

This dissertation is motivated by two trends influencing geoscientific activity and its subsequent societal impact: (1) the trend to increased production and use of digital geospatial data and related technologies, and (2) the greater undertaking of integrative science aimed at answering complex questions that require knowledge to be integrated within and between disciplines.

The first trend is a product of technological innovation, as advances in geospatial data management and analysis have coincided with the proliferation of digital geospatial data, while the emergence of the Internet has reduced technological barriers to the remote access of digital resources, such as data and tools, and incidentally increased the demand for both, creating in the process scientific, economic, and societal opportunities. For example, scientific opportunities are exemplified by the immense Earthscope project, which will deploy a variety of sensors across the continental US to monitor and map aspects of the crust and mantle (van der Vink, 2002). Of significance is the fact that Earthscope is predicated on having in place mechanisms to deal with volumes of digital data previously considered unmanageable. Likewise, economic opportunities are evident in the emerging marketplace for web-based digital data and related tools, as exemplified by the businesses built around services for geospatial location (e.g. GPS), navigation, web-based cartography, and others. Finally, societal opportunities are apparent in the efforts made by public agencies to increasingly translate both science and technology into units of societal relevance. A good example is the Canadian GEOIDE initiative, which partially justifies its geospatial research funding on the basis of its prospective relevance to decision-making (GEOIDE stands for “GEoMatics for Informed DEcisions”) (http://www.geoide.ulaval.ca/). The first trend thus highlights technological innovation in geospatial information, which coupled with improved access to digital resources, has helped spur scientific, societal and economic markets for geospatial data and related technologies.

The second trend is a product of the move to integrative science, in which pressing “big-science” problems, such as climate change, natural hazards, sustainable development, biodiversity, are being addressed through multi-disciplinary efforts. The key to resolving both the scientific and societal dimensions of these problems lies in the ability to leverage results within and between scientific disciplines, and finding credible ways of translating the scientific results into useful societal products.

A technological manifestation of this trend is the emergence of networks that connect data, tools, and people, i.e. “cyberinfrastructures” or “scientific grids”, whose expressed aim is to facilitate integrative science. Such networks currently exist for the geosciences, physical sciences, biosciences, ecosciences, and others (Hey & Trefethen, 2005), but in relative isolation to each other. However, this state of isolation is probably temporary, given the ample evidence that cross-disciplinary science is equally relying on integrative technology. For example, in the burgeoning field of integrated assessment modeling, modeling methods and technologies from the natural sciences, social sciences and economic disciplines are being coupled to produce holistic regional assessments for, e.g., sustainable development of cities or climate change (Parson & Fisher-Vandem, 1997; Robinson & Herbert, 2001). An implicit future direction can then be extrapolated from such crosscutting activity, one that
points to multi-disciplinary cyberinfrastructures as a foundation for integrative science. The second trend therefore highlights integrative science and its technological manifestation in the form of cyberinfrastructures, with the latter viewed as a new operational paradigm for achieving such science and readily delivering it in consumable form to other scientists and society.

1.2 Problem Statement: geospatial interoperability and the meaning gap

Underlying both motivating trends are some common requirements and issues. Both geospatial resource proliferation and integrative science require methods and tools for technological interoperability: i.e. interoperability of databases, software, instruments (including computers) and the people who operate them. Interoperability in this broad sense refers to the overall task of coupling such resources within and between cyberinfrastructures. It facilitates activities initiated by humans or machines for: data collection; finding appropriate data, tools and people for a given task; fusing data for input into tools that do modeling, analysis, visualization, etc.; chaining such tools within a designated workflow; and providing publication and deliberation services so that scientists can release results and deliberate amongst themselves about their validity and with others about their societal use. As the resources inherent in these activities are typically engineered independently, located in various geographic locations, and heterogeneous in their design and operation, interoperability efforts must aim to connect them into a loosely-coupled system. This implies the resources themselves are not fundamentally altered to operate within the system, but rather that connectivity is ensured through mediating hardware and software that play a bridging role between the resources and the system (Domenig & Dittrich, 1999; Wiederhold, 1996).

A significant barrier to such interoperability is the meaning gap (Goodchild, et al., 1999). At the core of this problem is the issue of digitally representing, transferring and understanding meaning, between and within human and machine agents. Because cyberinfrastructure is being developed as an open system that is persistently evolving, discovering what resources exist and deciphering what they are and how to use them is not readily apparent. One approach to solving this problem casts interoperability as a communication task, in which the goal is to understand messages passed between a source and target (Brodeur, et al., 2003), where both source and target can be some combination of data, tool and person. For example, a database must be able to communicate its content, structure, functions and operational interface so that appropriate transformations can be applied to enable ingestion of its data by another database, a tool or person; likewise a tool must be able to communicate its inputs, outputs, functions and interfaces so it can be appropriately invoked by human and other machine agents; and finally, people may wish to collaborate with each other as well as invoke tools and use data within cyberinfrastructure. In this sense, interoperability therefore requires aligning meanings of messages passed between, and within, human and machine agents (Pike, et al., 2004).

1.2.1 Bridging the meaning gap

If interoperability is viewed as a communication problem, as described above, then the issue of bridging the meaning gap reduces to the problem of producing and understanding a digital message. Moreover, if a message is considered to be a sign or collection of signs, then the study of digital signs, or computational semiotics, can provide a framework for the meaning
of messages. By way of following this general approach, the computational semiotic framework used in this dissertation is introduced here and discussed in detail in subsequent chapters (chapters 2 and 3). The framework is an amalgam of interdependent layers that are drawn from various sources (e.g. Klaus, 1974; Morris, 1938; Stamper, 1996), and consists of systems, syntax, schematics, sigmatics, semantics and pragmatics. The systems layer involves the protocols and physical infrastructure for passing messages. The syntax layer involves the grammar and lexicon used within a message. The schematics layer involves the structure of the document holding the message, and the sigmatics layer contains the actual message. The semantics layer involves descriptions of the concepts inherent in the message and schema, and pragmatics refers to contextual elements that frame how the message should be used, has been used and created.

For example, a message consisting of data from a database can be decomposed into these layers: the internet as a transmission media represents the systems layer; commitment to a mark-up language such as XML (Bray, et al., 2004) represents the syntax layer; commitment to a specific XML schema such as RDF-OWL (McGuinness & Harmelen, 2004), that defines primitives such as concepts for structuring the data, represents the schematics layer; packaging the message into the schema structure represents the sigmatics layer; commitment to a specific set of logic-based definitions for the concepts underlying the content and structure of the message represents the semantics layer; and the pragmatics layer contains information on the origins, effects and prior uses of the concepts and data, and can include contextual guidelines such as informal policies and protocols for how to use the concepts and related data. To be interoperable, a recipient would transform these layers into local versions without significant loss or alteration of meaning. For instance, if a database were to ingest the preceding message it would need to follow the same internet connectivity protocol, but reconstruct the message such that the syntax follows a database language, e.g. SQL (International Standards Organization, 2003), the structure of the message is transformed to fit the local database schema, and the content of the message is transformed into acceptable content for the database.

The example below demonstrates these transformations. In the example an RDF-OWL description of an instance of a geological unit is transformed into a SQL INSERT statement. The description in RDF-OWL is quite simple, consisting of a name for the concept, the “Baker Brook basalt”, and attributes for the higher order classification the unit is a part of (group_name), the class and type of rock (rock_class and rock_type), the geologic age of the unit (unit_age), and the geometric shape of the unit (poly_id).

<Baker Brook basalt rdf_id="unit_102">

<group_name rdf:resource="#Roberts Arm Group"/>
<rock_class rdf:resource="#Extrusive"/>
<rock_type rdf:resource="#mafic marine"/>
<unit_age rdf:resource="#Ordovician"/>
<poly_id rdf:resource="#poly_102"/>
</Baker_Brook_basalt>

Note that the SQL statement below displays syntactic, schematic, sigmatic and semantic differences to the RDF-OWL fragment above. Syntactically the message is now written in
SQL. Schematically, a table for units is assumed (GeologicUnit) and an attribute for the unit’s name has been added (unit_name). Sigmatically, the information differs in that the unit’s identifier “unit_102” is not required. Semantically (and sigmatically), the RDF rock_class value of “Extrusive” and the associated rock_type value of “mafic marine” lead to a conversion to the database’s convention of “volcanic, mafic marine” for rock_type, and the geometric shape of the unit has been converted into the database’s convention for storing polygons and assigned a numeric identifier of 102:

```
INSERT INTO GeologicUnit unit_name, group_name, rock_class, rock_type, unit_age, poly_id
VALUES "Baker Brook basalt","Roberts Arm Group", "Extrusive", "volcanic, mafic marine", "Ordovician",102
```

Although this example is quite simple, the problems associated with such transformations are in fact complex, with many unresolved problems especially in the geographical domain. A prime example is the issue of multiple representations where a geographic entity can be represented in multiple ways due to copying, versioning, evolution of real entity in time, as well as due to representational and scientific differences. Representational differences are evident when the geometry and location of a geographic entity vary according to geographic scale: a small region may be represented as a point at fine scales and a polygon at broad scales. Scientific differences are apparent when its geometry, location and classification vary according to scientific debate: e.g. two scientists might disagree about the borders of an entity, hence disagree about its geometry and location, as well as its classification (Bie & Beckett, 1973). Multiple classifications may also derive from perspectives that classify an entity according to its functional role (Fonseca, et al., 2002; Riedemann & Kuhn, 1999): e.g. a geologic unit can be viewed as a repository for nuclear waste, pollutants, water, etc.

A major consequence of multiple representations is the distribution of pertinent information amongst the various representations, where each representation potentially possesses varying degrees of accuracy and completeness. This implies that an agent who requests a specific geographical entity, say an instance of the ‘Baker Brook basalt’, from a distributed set of resources might then receive a multitude of messages with heterogeneous representations for the entity. While it might be possible to transform each message into the recipients’ layers, as shown above, the deeper question lies in how to select the most appropriate message, or how to integrate the relevant thematic and information distributed amongst the various messages. For, though it may be tolerable for information browsing purposes to receive and view multiple representations, it is often inappropriate for critical processing needs where a single item is required, e.g. disaster planners and responders generally prefer a single, albeit dynamic, description for a natural event such as a forest fire or earthquake.

1.2.2 Bridging the pragmatics gap

It is precisely in the answer to these deeper questions that pragmatics plays a role. Unlike traditional semantics which answers the questions “what is this entity, how can it be used?”; pragmatics answers the questions “why is this entity the way it is, and how has it been created and used?”. Pragmatics therefore speaks to the context, relevance and quality of an entity, which aids in the selection or integration of multiple representations. So, if an instance
of the ‘Baker Brook basalt’ concept is returned with two distinct descriptions, the knowledge that one is supported with age date analysis or more field sampling would likely cause it to be selected. Pragmatics thus helps evaluate an entity’s “fitness for use” (Chrisman, 1984).

The significance of evaluating fitness for use can be demonstrated via Figure 1.1, which shows a fragment of a geospatial database in which a single geographical region is classified into distinct geological units on adjacent sides of a map border (from Davenport, et al., 1999). The bottom pink unit is classified as the ‘Buchan’s Group’ concept and contains a mixture of rock types (not shown) whereas the upper blue unit is classified as the ‘Roberts Arm Group’ concept and contains a collection of sub-units (formations, beds, etc.), including instances of the ‘Baker Brook basalt’ concept. However, such a sharp break between adjacent units is highly implausible, given the absence of other significant geological factors such as faulting. Determining which representation to trust at the border, perhaps for some environmental decision such as nuclear waste disposal site selection, is then a matter for geo-pragmatics: semantics will inform what the differences are, whereas geo-pragmatics will help evaluate them by informing why and how those differences came to be, so the representations can be used appropriately and reliably.

![Figure 1.1: varying classification of geological regions across maps.](image)
But, what is the character of geo-pragmatics? Does the emphasis on explanatory context impact syntax, sigmatics, schematics and semantics—that is, given that meaning seems to involve explanatory context, how are the syntax, structure and content of geo-representations affected by such context, and to what degree? How does this mesh with geoscience, which often aims to explain the geoscientific history of geospatial regions—is geo-pragmatics somehow tied to such explanations... if so, how? These are some of the questions to be explored in this dissertation.

1.3 Objectives and Methods

Broadly speaking, the aim of this dissertation is to boost the meaning level of digitally represented geographical concepts by incorporating pragmatics into their representations. Because existing work on information pragmatics has revolved mainly around human-computer interface issues, viewing it from the perspective of natural or machine interactions is quite rare. To provide a holistic assessment, this dissertation takes a reconnaissance approach, exploring the impact of pragmatics on several aspects of representation that contribute to meaning, rather than focusing on any single component. Specifically, it explores the impact on the syntax, sigmatic, schematic and semantic aspects of representation, using examples from the geoscientific domain. It demonstrates that taking a geo-pragmatic viewpoint has significant ramifications for the organization of information and particularly concepts, potentially leading to improved interoperability of resources in cyberinfrastructures.

The key objectives are:

1. identify and characterize geo-pragmatic and semantic challenges to the interoperability of geoscience information;
2. formalize aspects of geo-pragmatics, to boost computational meaning levels;
3. evaluate geo-pragmatic artefacts in empirical data supporting some geoscience concepts;
4. explore the impact of geo-pragmatics on geoscience database schema;
5. explore the impact of geo-pragmatics on geoscience semantics.

The first objective aims to identify issues related to semantics and pragmatics in the interoperability of geoscience information. This requires an understanding of interoperability as it applies to the geoscience domain. To narrow the focus, the problem is constrained to the interoperability of databases holding information about geological regions, which involves related concepts (semantics) and their origins, effects and uses (geo-pragmatics). Challenges are identified through a critical analysis of the nature of these elements. The analysis maintains a conceptual tone, examining relevant philosophical, cognitive and computational literature and identifying gaps in representing geoscience knowledge in existing semantic frameworks. This objective is addressed in chapter 2.

The second objective aims to explore how pragmatic considerations in geoscience knowledge representation can increase meaning levels; it also aims to address some of the concerns raised in the first objective by developing a logic formalism and applying it to an example from geoscience. The formalism is first presented via a diagram showing mappings between core elements, such as concepts and their instances. Validity of the diagram is demonstrated
via a logic-based encoding of how a geological unit might be discovered, and success is measured by increases in the conceptual granularity of the representations that enable reporting on the origins of a represented entity. Only the relevant geoscience entities are encoded in the logic fragment, which acts as proof-of-concept, and no attempt is made to develop a complete logical system for this aspect of the domain. This second objective is addressed in chapter 3.

The third objective seeks empirical evidence for the presence of pragmatic aspects in geoscience activity, to demonstrate both their significance and the applicability of the results of the second objective. It analyses the development of concepts used to classify geographic regions, by inspecting the correlation of the observed field data with the concepts. In doing so, it tests the hypothesis that the concept’s originating situations, as represented by the field data, contribute to the understanding of the concept. Specifically, the field data collected by three geologists working together to develop a geologic map for a region are compared and contrasted over time. Significant differences in how the field data clusters over time is used as an indicator for the consistency and completeness of the concepts that they support, and as a reflection of the degree of impact of the origins of a concept on its meaning. Neural networks and statistical techniques are used to compare and contrast the field data. Success is gauged by using standard techniques for neural network assessment that compare test and trained data samples. This objective is addressed in chapter 4.

The fourth objective explores how a pragmatic viewpoint affects the design of database schema for geographical information. The critical issue here arises from the first two objectives: if concepts are discovered and evolved during geographical inquiry, as explored theoretically and empirically by the first two objectives, then database systems must be designed to store and manage dynamic concepts and their pragmatic origins. This represents a shift in database design inasmuch traditional approaches assume a static and predetermined suite of concepts upon which such database schemas are founded. More specifically the intention here is to develop a database schema for those geoscientific concepts used to classify geographical regions, taking into account their pragmatic origins and thus boosting their understandability. These types of concepts are typically described in the legend of a geological map, and are treated here as prototypical for many geographical sciences, such as soil science, ecology, and others. The success of the designed schema is demonstrated via example, by showing how the data from the previous objectives can be fitted into the schema, and practically, through implementation of the schema by external partners in their projects. This objective is addressed in chapter 5.

The fifth and final objective aims to explore how a geo-pragmatic viewpoint impacts the organization of concepts in ontology systems. The key issue here is similar to that in the last objective: approaches to organizing concepts in computational frameworks, called ontologies, are geared to representing either universal concepts, such as abstractions of patterns in nature, or subjective conceptualizations, such as mental contents, without taking into account their (geo-pragmatic) origin-dependence. So, this objective aims to develop an ontology organization that clearly identifies how origin-dependent concepts fit into ontology structures, and illustrates this organization by encoding a geoscientific example using existing knowledge representation techniques. Significant engineering obstacles are recognized but not addressed, as the main purpose is to develop a conceptual approach that enhances representation of the domain, leaving engineering issues to future work. Success is
therefore measured according to goodness of fit of the data to the conceptualization, via the example. This objective is addressed in chapter 6.

In summary, the overall approach of the dissertation is conceptual rather than engineering oriented or experimental. The main intention behind all objectives is to introduce and pursue a new conceptualization for geographical concepts and their representation, from a pragmatic perspective, thereby disambiguating several conflated aspects of geographical concepts. In meeting each objective, this dissertation introduces existing theoretical foundations and examines them from a pragmatic viewpoint, called geo-pragmatics, making adjustments if necessary, and testing those adjustments in lightweight prototypes. In order to cover the intended topics, the treatment of theory and development of prototypes is not comprehensive, but is instead quite focused and represents in many cases a first order glimpse of the issue. Because the dissertation deals with new theoretical and associated representation approaches, its success can be measured by goodness of fit of data to the representation, or in the case of the empirical component, goodness of fit of theory to data. Application of the results in existing representation mechanisms is demonstrated, but without deep consideration of the computational inefficiencies, which are recognized but not addressed. In general it is assumed that gains in conceptual clarity will eventually outweigh the added computational burden, which might in any case be overcome with engineering solutions. This dissertation will also not test the proposed benefits to interoperability, leaving this to future work.

This dissertation demonstrates several key results. It develops an improved understanding about the nature of geoscientific knowledge, including field data and its relation to geographical concepts, and the requirements for computing with such knowledge in cyberinfrastructure. It sketches a description of geo-pragmatics and demonstrates its representation with a logic-based encoding of a geoscientific example. It also develops a conceptual database schema suitable for organizing such data and concepts, improving geographical data organization, and demonstrates its viability via implementations by third parties, mainly governments and academia. In addition, a geo-pragmatic framework is developed for organizing geographical concepts and this framework is demonstrated using existing ontology languages, thereby broadening the overall understanding of geographical concept systems and their representation.

1.4 Outline of the dissertation

The dissertation is organized such that a single chapter addresses each objective. Moreover, the chapters are organized in building block fashion, as successive chapters utilize results of previous chapters. Because the substantive chapters cover different dimensions of the digital representation of the meaning associated with geographical entities, relevant existing results from prior chapters, as well as additional existing theory and methodology are introduced as required.

The chapters are organized as follows: existing approaches to geoscientific knowledge representation are covered (chapter 2), and some broad gaps related to geoscientific semantics and pragmatics are identified; these gaps are explored in subsequent chapters where a theory of geo-pragmatics is developed and demonstrated as a formalism (chapter 3); these theoretical results are empirically tested (chapter 4), and representational structures for databases (chapter 5) and concepts (chapter 6) are embellished in light of the theory and empirical evidence. The dissertation concludes with some notes on limitations of the work.
carried out, and with some indications of interesting future directions opened up by this work (chapter 7). References for cited work are provided next, and the Appendix that follows describes methodological details of the empirical study discussed in chapter 4.

1.4.1 Challenges to representing geoscience knowledge in cyberinfrastructure (chapter 2)
This chapter argues for the need to augment the representation of information ontologies with geo-pragmatic elements. Because of the conceptually disjoint nature of the geoscience information, interoperability systems need information ontologies to be built. Building information ontologies involves activities for their capture, representation and use, and these are generally constrained by the nature of geoscience knowledge and its requirement of representing context. Context is needed because geoscience knowledge is fragmented, has multiple perspectives and versions, and is often dependent on geospace-time regions. An important aspect of such contextualization is the scientific process of knowledge discovery and evolution, here called geo-pragmatics. The impact of geo-pragmatics on ontology representation is introduced in this chapter and explored further in the rest of this dissertation.

1.4.2 Representing geo-pragmatics (chapter 3)
This chapter lays the theoretical groundwork that will be empirically tested and computationally applied in subsequent chapters. It introduces the pragmatic aspects of geoscientific representation, i.e. geo-pragmatics, develops a formalism and applies it to a realistic example from geoscience using logic. Geo-pragmatics is discussed in terms of a knowledge cycle, adapted from AI, in which geo-pragmatics fills the role of an explanation device—it explains how the components originate, what they affect and how they are used. A machine-driven view, one of cyberinfrastructure, is generally adopted in which it is assumed that all components of the knowledge cycle can be represented to some degree in networked machines, resulting in the recognition of the machine as well as human and natural dimensions of geo-pragmatics.

1.4.3 Geoscience syntactics: experiments to illustrate the geo-pragmatic nature of geoscientific information (chapter 4)
This chapter tests the hypothesis that pragmatic artefacts are present in the information collected by some geoscientists collaboratively mapping a region. The field data of three geologists are compared and contrasted using geocomputational methods. These geologists worked as a team to develop a common map and a suite of associated concepts used to classify geospatial partitions of the map. Differences between and within individuals’ data for the same area would signify varied understanding of the concepts, and signal a need to better capture the pragmatic situations in which the concepts were created and applied. Results are presented which corroborate this hypothesis, suggesting such concepts are influenced by natural, human, theoretical, and situational factors.

1.4.4 Geoscience Schematics: structuring GIS databases for geo-pragmatics (chapter 5)
In this chapter a conceptual schema is developed for representing geo-pragmatic elements. Its intended application is database design, though it could hypothetically be adapted for other
systems such as ontologies. The schema directly models both the origin and some effects of geoscience concepts and instances, resulting in these elements being directly included in the database contents. This represents an abstraction level above the level typical for database representation. The shift is due to the fact that in many geoscientific disciplines concepts are regularly discovered, evolved and versioned, but the architecture of geospatial information systems is primarily aimed at supporting static conceptual structures in which concepts are predetermined and do not change in step with scientific progress. The result is a gap, addressed in this chapter, between our evolving understanding of these concepts and how they are represented in our systems. The schema is represented as a UML diagram. Its prototype implementation by both the author and third parties is also summarized in this chapter.

1.4.5 Geoscience Semantics: levels of abstraction in geoscience ontologies (chapter 6)

This chapter develops a framework, called LEKXIS, for organizing geographical concepts into six levels of abstraction. LEKXIS is derived from a synthesis of distinctions found to be similar across many fields: philosophical kinds, cognitive categories, and cognitive semantics. Discussed are these roots, as well as the resultant framework, which is demonstrated using an example from geoscience. Ontology implementation is demonstrated via representation in two popular ontology syntaxes: UML and OWL, and some potential applications are discussed.

1.4.6 Conclusions, Limitations and Future Directions (chapter 7)

This dissertation concludes with the notion that a pragmatic dimension to geoscientific meaning representation is not only realistic, but also viable and valuable. It does, however, face significant computational challenges to fully operationalize; many of the challenges are recognized but not dealt with. Apart from the outstanding computational challenges, remaining theoretical challenges are also identified.

1.5 Contributions to GIScience

The main contribution to GIScience of this dissertation rests in its introduction of pragmatics to geographical knowledge representation, i.e. in geo-pragmatics. In particular, geographical concepts have recently received considerable attention in GIScience, mainly under the rubric of ontologies (for a survey see Agarwal, 2005) and primarily from a semantic perspective where they are characterized in terms of:

- inheritance from fundamental philosophical categories, e.g. ‘Baker Brook basalt’ derives from ‘Geographical Region’ (Grenon & Smith, 2004);
- specification of necessary and sufficient logical conditions, e.g. ‘Baker Brook basalt’ is necessarily composed of extrusive rocks (Visser, et al., 2003);
- the potential role of concepts in theories, including scientific theories, e.g. in a specific theory the ‘Baker Brook basalt’ might possess certain attributes but not others (Fonseca, et al., 2002);
- similarity of instances, where instances are variously typical of concepts, e.g. the similarity between instances of the ‘Baker Brook basalt’ (Smith & Mark, 1998);
the potential functions that concepts could perform and the effects those functions
could entail (Riedemann & Kuhn, 1999; Kuhn, 2001), e.g. the function of the ‘Baker
Brook basalt’ in local geological processes;

the role of concepts in overall space-time representation (Peuquet, 2002).

Pragmatic aspects are largely untouched by these approaches. The dissertation therefore fills
a noticeable gap in understanding the character of geographical concepts and their
organization in computers mainly through inclusion of their pragmatic origins, e.g. where the
meaning of ‘Baker Brook basalt’ involves its evolutionary history and local geospace-time
situation, as well as the scientific processes and human actions involved.

1.6 Existing peer-reviewed publications arising from this dissertation

Segments of this dissertation are reproduced from the following peer-reviewed publications
with permission from the publisher: portions of chapter 2 are drawn from [1]; portions of
chapter 4 are from [3], [6] and [8]; and selected portions of chapter 5 are from [4] and [5].
Only original work carried out by the author is included in this dissertation without additional
citation. Shared work is explicitly identified and a citation provided.

Cyberinfrastructure: challenges, approaches and implementations. Geological Society of
America, Special Volume on Geoinformatics.

Scientific Data: From Data Integration to Scientific Workflows. Geological Society of
America Special Volume on Geoinformatics.

computing geological categories from field data. Computers & Geosciences, 30(7):719-
740.

(Eds.), Geographic Information Science-2nd Int’l Conference, GIScience 2002, LNCS

Richardson, D., van Oosterom, P. (Eds.), Advances in Spatial Data Handling, 10th

for geographic knowledge representation. In: Proceedings of the Ninth ACM
International Symposium on Advances in GIS, Atlanta, GA, Nov. 9-10, 2001. ACM

knowledge construction: filling in the gaps between exploration and explanation. In:
Richardson, D., van Oosterom, P. (Eds.), Advances in Spatial Data Handling, 10th

of geographical categories. In: Proceedings, 5th International Symposium on Spatial
Accuracy Assessment in Natural Resources and Environmental Sciences. July 10-12,
Melbourne, AU.

*With kind permission of Springer Science and Business Media
CHAPTER 2: CHALLENGES TO REPRESENTING GEOSCIENCE KNOWLEDGE IN CYBERINFRASTRUCTURE

This chapter suggests that collections of concepts, called information ontologies, need to be augmented with geo-pragmatic elements. The argument proceeds as indicated in Figure 2.1: because of the conceptually disjoint nature of the geoscience information, interoperability systems need information ontologies to be built. Building information ontologies involves activities for their capture, representation and use, and these are generally constrained by the nature of geoscience knowledge and its requirement for representing context. Context is needed because geoscience knowledge is fragmented, has multiple perspectives and versions, and is often dependent on geospace-time regions. An important aspect of such contextualization is the scientific process of discovery and evolution behind some knowledge, here called geo-pragmatics. The impact of geo-pragmatics on ontology representation is introduced in this chapter and explored further in the rest of the dissertation.

Figure 2.1: outline of chapter 2.

The sequence of the chapter follows the argument: the state of geoscience information and approaches to its interoperability are briefly summarized, and the need for ontologies is established (sections 2.1-2.2); the general nature of information ontologies is explored next, and some challenges to their engineering are identified (section 2.3). Geoscience ontologies are then explored by examining the distinct nature of geoscience knowledge, and further challenges are identified, amongst them the need for ontology representations to accommodate context (section 2.4). Geo-pragmatics is then introduced and explored as a means of achieving context, at least in part, raising issues that are partially addressed in the remainder of the dissertation (2.5). By way of closing the chapter, major assumptions and challenges are reviewed in section 2.6.

2.1 Geoscience information

Geoscience information is important to many sectors of society including science, business and government. Increasingly, these sectors’ activities are being coordinated to resolve pressing societal concerns such as climate change, land use, natural hazards, biodiversity, resource development, navigation and positioning, etc (Applegate & Bartlett, 2000). Achieving coordination involves increased interaction within sectors as well as between them, in that, for example, scientists must frequently exchange, share and integrate knowledge between themselves and with government and industry.

However, use of geoscience information is inhibited by its fragmentation along agency and thematic boundaries: the information is resident in autonomous databases that for historical reasons have been developed and maintained independently.
reasons are hosted in a variety of sectors, including government, industry and academia, and that are segmented thematically as per stratigraphy, geochronology, geophysics, etc. Expected increases in both information volume and web-based accessibility suggest such fragmentation will continue to rise, exacerbating difficulties in finding, integrating and ultimately using the information. This leads to a scenario in which routine information handling and research are seriously impeded, particularly in light of the prevailing trend in all sectors toward large-scale, multi-disciplinary projects that are increasingly reliant on computational tools and digital information. In response, several initiatives are pursuing greater cohesion of information through database networks in which multiple source databases are linked to provide network-wide access to information and associated tools (e.g. GEON, www.geongrid.org; SCEC, www.scec.org/cme/; CGKN, www.cgkn.net; NGMDB, http://ncgmp.usgs.gov/ngmdbproject/).

Database network initiatives such as these must overcome many thorny technological hurdles, including those associated with the following tasks: (1) registering information and tools with the network—the resources registration task (2) finding relevant information and services across the network—the resources discovery task; (3) integrating the discovered information—the information interoperability task; (4) integrating the various software tools—the services interoperability task; (5) using the integrated information and tools in scientific workflows—the (possibly collaborative) analysis and processing task; and (6) disseminating the scientific results, derived from the collective information, for scientific or educational purposes—the publishing task. This chapter primarily considers information interoperability (task 3), mainly as a means of identifying several challenges to the overall problem of interoperability of geoscientific information (or geoinformation).

2.2 Information Interoperability

Sharing information in database networks requires not only adequate geoinformation transportation, or the lossless transfer of data (Kuhn, 1997), but it further requires the lossless transfer of meaning (Goodchild, 1999) so that geoinformation can be well understood and appropriately and reliably used. The study of the mechanisms for achieving such lossless transfer can be called information interoperability.

In this dissertation information interoperability (henceforth “interoperability”) will specifically refer to the ability of a human or machine agent to: (1) receive unified and integrated results in response to a query of a database network, and to (2) update the contents of the network, such that the updates are available to users of the network. An example of the former, querying, is a request to the network for all mapped geological formations located in a geographic region of North America that contain a certain rock type and age—without having to individually search each source database, such as those for geological maps, lithostratigraphic lexicons, geochronologic databases, etc., and without necessarily having to reconcile differences in format and content manually. An example of the latter, updating, involves adding or deleting information by posing a single request to the network that is then executed within the appropriate source databases.

Interoperability is difficult because of distribution, autonomy and heterogeneity (Sheth & Larson, 1990). Distribution refers to how the information is spread amongst the multiple databases and their physical locations. Autonomy refers to the independence of each database within the network, including its control over local database design, response to external
requests, and policies for sharing information. Heterogeneity refers to differences between databases that can be systemic, syntactic, schematic, and semantic (Sheth, 1999): i.e. the databases can be implemented on different computing platforms (systems), use different control languages and file formats (syntactics), be diversely structured (schematics), possess different content, have diverse concepts for the same content and even have the same concepts for diverse content (semantics). To achieve interoperability, strategies for distribution, autonomy and heterogeneity must be implemented: the databases need to be connected and the information appropriately distributed; the database stewards must agree how to share some information and functions, and decide how to manage changes and transactions; and mechanisms must be adopted for unifying the information along the four dimensions of heterogeneity. Figure 2.2 illustrates reconciling information in these dimensions to enable exchange of meaning (adapted from Bishr, 1998; Gahegan, 1996). Note that the sigmatics layer introduced in chapter 1 is not explicitly recognized in interoperability strategies, but is encompassed within the schematic or semantic layers.

Also note that Figure 2.2 also extends the conventional framework with the addition of an additional layer called pragmatics. As much as the semantic layer is ontologic, dealing with already formed domain concepts, the pragmatic layer is equally epistemic, dealing with the origins leading to the formation of those concepts, as well as to its actual uses and effects. The pragmatic layer thus details how and why a concept is as it is, to enable scientific evaluation of concepts and promote their evolution, whereas the semantic layer details what the concept is. The notion that meaning has a pragmatic dimension of considerable weight is argued compellingly by philosophic, computational and geoscientific semioticians (Sowa, 2000; Baker, 1999). Both geoscience semantics and pragmatics (i.e. geo-pragmatics) are discussed further below.

In terms of database networks, the distribution, autonomy and multi-level heterogeneity requirements lead to a variety of possible database network architectures based on the level of coupling between the source databases and the overall database network (after
Elmagarmid, et al., 1999; Sheth & Larson, 1990). The discussion of these architectures below focuses mainly on semantic heterogeneity, categorizing the architectures according to the degree of overlap of the concepts in the source database schema and contents: i.e. including concepts in the database structure (i.e. in entities, relations, attributes, rules, functions) and in the vocabulary of the database contents (i.e. in attribute ranges, metadata).

- **Tightly Coupled** systems are characterized by low autonomy, information centralization and low heterogeneity (high conceptual overlap), enabling the implementation of a common database schema across the various source databases (Hull, 1997); e.g. a data warehouse in which information from the source databases are duplicated (i.e. materialized) and where a common schema, or some subset of it, is used by each source database and where mappings between the global and sources are minor because the source and global schemas are essentially the same.

- **Loosely Coupled** systems are characterized by moderate autonomy, distribution and heterogeneity amongst the sources. Partial conceptual overlap enables the creation of a common, single global schema for the network, built on shared concepts and the union of non-overlapping concepts. The global schema is thus somewhat independent of the source schema, requiring mappings to be developed between the global and various source schemas to enable information integration. These mappings are typically implemented as logical views between the global schema and the source schema: e.g. the global schema can be defined as a global view over the source schema (GAV: global-as-view), the local source schema can be defined as views over the global schema (LAV: local-as-view) (e.g. Ullman, 1997), or some hybrid (Cali et al., 2002; McBrien & Poulouvassilis, 2003). The global schema typically serves as a central conceptual portal into the database network, providing the illusion of a single unified data source.

- **Disjoint** systems are highly autonomous, distributed and semantically heterogeneous, precluding the creation of a global schema based primarily on source concepts. Interoperability is achieved by mapping source concepts onto more general knowledge, an ontology, consisting of general concepts, theories, models, processes, and other knowledge forms. Such an ontology can serve many roles in a disjoint system, providing: (1) metadata about the network resources; (2) a global conceptual schema for the network that connects both data and tools; (3) a vocabulary for the network contents; (4) the foundation for a query interface that can act as a central portal into the network; (5) a representation of geoscientific knowledge. These roles are variously combined or compartmentalized within different implementations (Wache et al., 2001). When disjointness is extreme, perhaps because very different types of information sources are regularly added and deleted from the network, mapping to a common domain ontology may be unachievable and information might be integrated via peer-to-peer translations between sources (e.g. Fensel, et al., 2001; Gribble, et al., 2001; Panti et al., 2002).

As implied by Table 2.1, the boundaries between these architectures are fuzzy, reflecting a gradient in the overlap of source concepts.
Table 2.1: characterization of architectures for database networks.

<table>
<thead>
<tr>
<th>Autonomy</th>
<th>Distribution</th>
<th>Heterogeneity</th>
<th>Common Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tight</td>
<td>low</td>
<td>central</td>
<td>low</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loose</td>
<td>high</td>
<td>distributed</td>
<td>high</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disjoint</td>
<td>high</td>
<td>distributed</td>
<td>high</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The marked prevalence of semantic heterogeneity in geoscientific databases suggests that the adoption of disjoint architectures might be the most applicable approach, and this consequently gives rise to the need to develop a shared ontology from the conventions of geoscience. This semantic focus would supplement existing syntactic (e.g. Cox, 2001; Ramachandran et al., 2001) and schematic (e.g. Breunig, 1999; Gartner et al., 2001) interoperability orientations in geoscience. But, what is the character of this semantic focus and how does it apply to geoscience? In addition, will such a semantic focus sufficiently enable interoperability? This dissertation takes the position that a semantic focus is insufficient and that interoperability solutions also require a pragmatic dimension to properly contextualize concepts. But how does pragmatics apply to geo-information? The next three sections explore these questions. Section 2.3 summarizes approaches to ontologies and identifies some challenges for its use in geoscientific interoperability; section 2.4 explores the nature of geoscience knowledge and its representation in ontologies; and section 2.5 explores the pragmatic perspective on ontologies for enhanced representation.

2.3 Ontologies and information interoperability

The ontologic approach to knowledge representation assigns meaning to information based on the position of the related concept within a network of concepts within some theory. It is discussed below in terms of philosophic ontology and information ontology.

2.3.1 Philosophical Ontology

As described by Sowa (2000, 51-67), philosophical ontology is concerned with uncovering ‘true’ and universal categories, i.e. philosophers such as Plato, Aristotle, Kant, Peirce, Husserl, Whitehead and Heidegger identified categories such as quantity, quality, relation, space, time, etc. Of interest is the fact that later categories (such as Peirce’s, Husserl’s and Heidegger’s) are human-centered, in that they form a basis for meaning that relies on how humans interact with their surroundings: for Peirce meaning accrues as humans respond to internal or external signs, for Husserl it is a product of human intentionality, and for Heidegger it is generated by culture. Universally true categories are referred to as ‘top-level’ or ‘upper-level’ and are generally uncovered ‘top-down’ through reasoning and logic, and are
arranged in taxonomic structures such as lattices (ibid, 96-109). They are obviously significant to meaning representation as they form a basis for the derivation of less generic categories, such as those developed by geoscientists. Meaning in philosophic ontology is thus related to the underlying essence of the universal categories and their organization.

2.3.2 Information ontology

Information ontology (henceforth “ontology”) is concerned with how categories and concepts, such as scientific concepts, are represented for computation (Gruber, 1995). It is defined as an explicit specification of shared concepts and theories (Gruber, 1993), one that represents the intended meaning of a vocabulary (Guarino, 1998). Such ontologies can in general be seen as having two purposes, enabling computer operations as well as aiding human comprehension (Fensel, et al., 2003). In the former, an ontology can provide formal and explicit (value-free) meaning for information, facilitating computer-based information processing; in the latter, an ontology can stimulate informal and possibly implicit (value-laden) meaning in humans to enable understanding of information and theory.

The computational focus of information ontology distinguishes it from philosophic ontology, as the former tends to be more concerned with digitally representing mental concepts than discovering a priori truths (Guarino, 1998; Smith, 1995; 1998a,b); however, overlap occurs when philosophic inquiry is used to identify the elements fundamental to digital knowledge representation (e.g. Guarino, 1995; Grenon & Smith, 2004). Then philosophical invariants are adapted for computation through their logical encoding: examples of such invariants include the distinction between entities that persist in space-time (endurants: material objects) and those that exist in time (perdurants: processes); likewise the distinction between a property (quality) and the object to which it inheres, and others. For example, the DOLCE ontology in Figure 2.3 uses these distinctions (from Gangemi, et al., 2003).

![Figure 2.3: the DOLCE upper level ontology (from Gangemi, et al., 2003).](image-url)
Apart from philosophical invariants, cognitive invariants are also being implemented as the foundation for information ontologies. This involves the grounding of spatial predicates in primitives thought to be built up from common mental hardware and sensory exposure to spatial situations (e.g. ‘on’: Frank & Raubal, 1999; Probst & Lutz, 2004).

Regardless of their philosophical or cognitive foundations, upper-level ontologies are intended to serve as the foundation for domain ontologies: e.g., geoscientific endurants could be specialized into rock materials, geological formations, etc., which would obey the high-level philosophical invariants. An upper-level ontology based on some coherently related suite of invariants would therefore seem to be a requirement for geoscience ontology. However, this does not necessarily imply that one upper-level ontology should be privileged. Instead, for this dissertation, it is only assumed that domain ontologies would benefit from some coherent upper-level framework that would help guide ontology design for a specific purpose. Moreover, this dissertation will argue in subsequent chapters for the inclusion of supplementary meta-functions and meta-concepts in such a framework.

### 2.3.3 Types of information ontologies

Information ontologies are often categorized according to content or representational structure. Content-based approaches commonly distinguish between ontologies that apply across domains and those that pertain within domains; e.g. formal ontologies represent general a priori concepts such as ‘relation’, whereas domain ontologies represent domain concepts such as ‘fault’. When the domain entities are fundamentally material, as in the geosciences, the domain ontology is called a material ontology (Smith, 1998a). Structure-based categorizations distinguish ontologies according to the mode of representation and the nature and degree of connectivity between entities in the ontology (e.g. McGuinness, 2003). An important categorization based on these factors occurs in (Sowa, 2000, 493-496) where linguistic, axiomatized and prototypical ontologies are differentiated. Linguistic ontologies consist of loosely-structured standard terms and their natural language definitions, such as a list of terms in a glossary (e.g. Jackson, 1997), or they are exemplified by more structured entities such as thesauri that include term relations, e.g. ‘synonym’. Most notably, linguistic ontologies are difficult to compute with but eminently comprehensible to humans. In contrast, axiomatized ontologies are quite computable as they use a logical language to represent the concepts and relations that underlie terms (e.g. Karp, 2000; NADM, 2004), but the computation is at present limited to classification issues that are well understood by logical systems, such as verifying the logical consistency of taxonomic relations. In contrast to both linguistic and axiomatized ontologies, concept meanings in prototype ontologies are not defined a priori but are derived from prototypical examples, e.g. a geological formation described in terms of its type locality(ies). Prototypical concept definitions may not therefore be exact or crisp, but may instead be approximate summaries of formative evidence. In this way they explicitly recognize the pragmatic origins of such concepts.

From a structural viewpoint, then, material geoscience ontologies may be linguistic or a mixed ontology containing axiomatized as well as prototypical concepts. The relationship between a linguistic and a mixed ontology in a database network involves connecting terms and axiomatic-prototypical representations of term meanings. For example, synonymy, where multiple terms have a common meaning, can be viewed as the mapping of multiple linguistic terms onto a single axiomatic and/or prototypical concept, e.g. the common concept for “strike” and “azimuth”; and polysemy, where a single term has multiple meanings, can be
viewed as the mapping of multiple axiomatic and/or prototypical concepts onto a single term, e.g. a geological unit name that refers to both a geological formation and a member.

2.3.4 How can ontologies be used for geo-information interoperability?
A disjoint architecture for geoscience interoperability could then benefit from: (1) a mixed material ontology, to bridge disjoint source database concepts and terms as well as link to prototypical evidence, and (2) a linguistic ontology that catalogs standard terms and maps them onto concepts in the axiomatic ontology (Klein, 2002). This architecture suggests at least four types of mappings, as shown in Figure 2.4.

![Figure 2.4: integrating geoscience map databases using ontologies.](image)

The four mappings occur: (1) between the source databases and the mixed ontology to deal with synonymy and polysemy in the database contents and schema (dashed lines, right); (2) between concepts and the prototypical evidence in the source databases, to document the (pragmatic) origin of concepts (solid line, right); (3) between the mixed ontology and the linguistic ontology to identify standard vocabulary for concepts (solid lines, middle); and (4) between the ontologies and the user interface to specify requests and communicate data in
one terminology to people, tools and other query sources (dashed lines, left). The number of the mappings implemented will increase in a pluralistic system in which multiple linguistic and mixed ontologies are supported. The justification for a pluralistic approach rests in the notion that the task of science is to evolve knowledge, implying that ontologies will likely evolve through time and via conflicts, following scientific progress and debate.

2.3.5 Challenges to ontology-based interoperability

General challenges associated with ontologies can be grouped into issues that deal with ontology capture, representation, and use. Ontology capture involves the determination of the contents of an ontology, typically using philosophical, cognitive, social or computational analysis techniques. Once discovered, the contents require representation in some formal or informal system, and once represented the ontology can be used, though use is constrained by both the scope of the contents and the type of representation.

Some challenges to capturing ontologies for geo-interoperability

It is important here to distinguish knowledge discovery (or recognition) from knowledge capture, in that capture refers here to the process of moving established knowledge into a computable representation. Discovery must then precede capture, unless the discovery occurs in a computational environment itself (e.g. Langley, 2000), in which case the two processes conflate.

A starting point for ontology capture might be to mine the concepts existing implicitly and explicitly in database schemas. But database schema and ontologies are not equivalent. A database schema and ontology can be differentiated on the basis of purpose as well as conceptual and structural completeness: database schema (conceptual, logical and physical) are geared to the specific requirements of a suite of data, particular tasks, and system constraints, and are not typically fully axiomatized, whereas ontologies are application independent and possess broader, more general and axiomatized content (Fonseca, et al., 2003). Hence migration of database schemas to ontologies requires generalization and axiomitization of the schemas. This further requires a social process of reaching consensus on the meaning of the database contents (Harvey, 1999), and agreeing which concepts should directly migrate to the ontology and which should be generalized. Associated with this is the problem of capturing tacit knowledge resident in people’s heads, which seems to call for cognitive or psychological capture methods. Other sources for ontology contents are textbooks and classification schemes that, although scientifically useful, are not often stated in language that is easily translated to rigorous computable form; this calls for methods to capture concepts from scientific and other documents. Finally, capturing the most general types of concepts such as those in the upper level is typically accomplished via philosophical analysis (e.g. Guarino & Welty, 2002).

Ontology challenge 1: Developing and using social, cognitive, lexical, and philosophical methods for capturing ontologies inherent in groups, heads, documents and the world.

Some challenges to representing ontologies for geo-interoperability

A major question to be answered by any system designer implementing ontologies is which ontology to implement? That is, will the system privilege one ontology, or provide several, and if the latter will it facilitate mappings between them? The representational issue raised
here is one that deals with perspectives and versions—will the ontology structure enable concepts to be viewed from multiple perspectives and allow concepts to evolve or to compete in any one perspective? Perspectives here involve the various roles a concept can play (Fonseca, et al., 2002). These roles typically contain disjoint properties such that perspectives contain different descriptions of a concept, e.g. the concept of ‘native land claim’ may be variously described by those making the claim, those administering the claim process, and those opposing the claim. Versions are defined here to occur when a concept’s description changes in time within a perspective, or when multiple competing views are held for a concept within a perspective. Of note is that scientific concepts exemplify both conceptual perspectives and versions: a concept for a geological unit can have properties obtained from fieldwork, such as color, thickness, etc., and properties obtained from remote-sensing such as a characteristic spectral signature. Both of these can evolve with more collected data or with physical changes to the concept’s instances, and both can be subject to alternate opinions due to varying methodology, assumptions, motivations, etc.

Though perspectives and versions enable concepts to be described variously, they do not explain the concept. Explanation is widely to considered to require greater context, including the lineage and effects of concepts. Lineage is particularly important to scientists in that it enables concepts to be validated by other scientists. Finally, concepts are not always crisp, but may be graded and have indeterminate boundaries, requiring appropriate methods of representation.

**Ontology challenge 2: Representing perspectives, versions, contexts and graded concepts.**

**Some challenges to using ontologies for geo-interoperability**

Ontology use refers to the deployment of an ontology in a computational environment, after it has been captured and represented. Effective use of ontologies is currently hindered by the lack of applications that deploy them, and the lack of ontologies available. Although this situation is likely to change given the resources being expended on ontology projects such as the Semantic Web, it is a current reality. From the viewpoint of interoperability, some critical questions need to be asked, and eventually answered: how much automated integration is practical? When should humans make decisions instead of machines? At what point are ontologies too impoverished to accomplish the desired task, and require human intervention? The integration of scientific information is a case in point: information integration in science often involves steps in a hypothesis building or testing process, and complete reliance on machines will be realized only once automated scientific discovery is achievable. Until then, interoperability will likely be a patchwork affair in which established and constrained scientific processes will utilize ontologies in the background for automated information integration, but where humans will likely drive innovative exploratory work (e.g. Ludaescher, et al., in press). For example, information integration steps might include analytic, simulation, and other scientific processes in a workflow. Interoperability thus extends beyond simple database integration, but reaches into the domain of complex science activities that are underpinned by cyberinfrastructures.

**Ontology challenge 3: Developing ontology-enabled workflows from scientific activities.**
2.4 Ontologies and Geoscience Knowledge

What constraints does geoscience knowledge impose on ontology capture, use and representation? Do geoscience ontologies fit the mould described above for ontologies in general? A careful look at the nature of geoscience knowledge is required in order to approach these questions.

2.4.1 Perspectives on geoscience knowledge

Knowledge, including scientific knowledge, has historically been approached in many ways. These ways might be roughly lumped into three groups distinguished by their positions on the origins and location of knowledge. The first position suggests knowledge resides in external reality independent of humans but variably accessible to them (i.e. philosophical realists). The second position suggests that knowledge is human-dependent and is a function of cognitive and/or social action (philosophical mentalists/constructivists). The third position holds that knowledge resides in language and is a function of linguistic foundations and actions (i.e. philosophical nominalists).

The main debate amongst geoscientists has occurred between realists and constructivists: scientific realists hold that increasingly universal knowledge about an objective biophysical reality is obtainable via scientific method, while constructivists argue that knowledge is fundamentally relative to the social forces experienced by the individual, and middle positions suggest that approximate and incremental knowledge of biophysical reality is methodologically obtainable by individuals participating within an affecting social milieu (Gould, 2000). These commitments have not only philosophical relevance, but also have critical impact on the representation of knowledge (Raper, 1999; 2000). A strong realist stance implies that knowledge can be directly identified and represented, that a single ontology for geoscience is obtainable and representable, that automated machine integration of data is feasible, and that pragmatics play a minor role in representing meaning. A strong constructivist stance implies that knowledge is relative, that cognitive and social contexts should be represented, i.e. their pragmatics, and that satisfactory data integration must be performed ultimately by humans but will be nonetheless inherently imperfect. Lastly, intermediate positions emphasize the approximate and evolving nature of knowledge, implying that multiple ontologies and geoscientific interpretations of varying certainty should be represented and that machine-based and/or human-based data integration may be required depending on certainties, complexities, etc., in the knowledge, methods and data—this position also needs pragmatic aspects to be represented.

Underpinning these positions are differing sentiments about how the knowledge acquisition and evaluation processes (pragmatics) impact the identified knowledge (semantics). Recent commentary among geoscientists adopts the intermediate position in which geoscience knowledge is variably bound to the discovery process (Baker, 1996; 1999; Frodeman, 1995; 2003). Investigating these characteristics provides insight into how such knowledge should be represented.

2.4.2 Realist perspective: unique characteristics of geoscience knowledge

The realist view of geoscience knowledge emphasizes natural pragmatics, i.e. the cause-effect relations occurring in nature between entities. In this view, geoscientific knowledge differs from knowledge in other sciences in terms of both natural characteristics and the
limitations in scientific process for their discovery. These differences are stated here in terms of complexity, incompleteness, indeterminacy, singularity, explanation, historicity, and generality:

- **Complexity**: geological knowledge is complex because its subject, the physical Earth system, is complex, both in terms of heterogeneity in the composition of materials and in terms of the multiple processes interacting with those materials at various spatial and temporal scales. There are therefore a large number of variables (concepts) and interdependencies involved in geoscientific knowledge.

- **Incompleteness**: geological knowledge is incomplete due to under-sampling of the complex physical Earth; the sampling itself is limited by instrument, spatial, temporal or socio-economic constraints. This implies geological knowledge forms an open system that evolves with additional observation. As all the relevant variables may therefore not be identified, there exist gaps in both knowledge and data (Mann, 1993). Geoscience knowledge is therefore bootstrapped from an imperfect knowledge base.

- **Indeterminacy**: complexity and incompleteness imply understanding is fragmentary, in that the vastness and complexity of the variable space cannot be uniquely solved by available data, leading to multiple competing models to describe a single situation that are all under-determined in varying degrees. Hence geological knowledge can be variously interpreted and is inherently uncertain (Schumm, 1991).

- **Singularity**: because of the preceding characteristics geological situations are on the whole quite differentiated and thus only partially transportable, inhibiting their complete reproduction in laboratory situations and limiting the generality of findings. However, specific elements are transferable and might be studied in isolation, such as foundational elements of the theoretical framework. (e.g. laws, processes; Frodeman, 1995).

- **Explanation**: geoscience is more often aimed at explaining a specific physical situation, by describing characteristics and interpreting them in relation to background concepts and theory (e.g. plate tectonics), than generalizing situations into empirical regularities, new concepts or theories, or predicting future situations. The primary product is thus an explanatory model: a reconstruction of the state of earth materials and processes, at some spatio-temporal location, and their causal interactions with historical materials and processes. Due to indeterminacy and singularity, the criteria for selecting the optimum model is its overall coherence with observations and background knowledge rather than experimental replication. Coherence is often conveyed via graphic and textual narratives such as maps and reports using scientific and rhetorical reasoning (Engelhardt & Zimmermann, 1982).

- **Historicity**: the role of history in explanation is emphasized in two respects: (1) both the objects to be explained as well as their causes can occur in the past, providing geoscience with a largely causal imperative, and (2) explanations often involve longer time periods than in other sciences.

- **Generality**: geoscience knowledge can be grouped into two broad categories: knowledge that might apply to any geospace-time situation such as the ‘fault’ or ‘process’ concepts, and knowledge that can apply to only one such situation such as the situated concept ‘Baker Brook basalt’. In this dissertation, the former is called theoretical (\( T_s \)) and the
latter is called situational ($S_k$). Explanatory models can then be said to consist of situational knowledge, whereas more abstract theories can then be said to consist of theoretical knowledge.

These insights particularly reinforce the representation and use challenges above: plurality in geoscientific ontology is needed to account for multiple explanations, theories, and knowledge evolution, and integration mechanisms are needed to combine the plural or fragmentary views and map between them.

2.4.3 Constructivist perspective: social-cognitive aspects of geoscience knowledge

Socio-cognitive aspects of geoscientific knowledge emphasize the human function in obtaining geoscience knowledge (Arbib & Hesse, 1963) thus highlighting human pragmatics. These aspects build on the categories for situational and theoretical knowledge suggested above.

The first important distinction is cognitive, where implicit ($I_k$) and explicit ($E_k$) knowledge are distinguished (Goschke, 1997). Implicit knowledge refers to knowledge that is not expressible but is tacit, intuitive, or procedural, e.g. one can be told how to ride a bicycle, but full knowledge rests in the experience of riding. It can be seen as applying to a person as well as to a group. Implicit knowledge may, in a limited sense, be conveyed via loose explicit structures such as narratives and via more formal structures that represent situations and contexts. It is important to geoscience because of well-recognized ability of geoscientists to develop holistic understandings about the natural environment through direct experience, i.e. fieldwork (Baker, 1999; Frodeman, 2003). Explicit knowledge, on the other hand, is readily expressed in some language, is often grounded in rational and conscious thought, and can apply to individuals or groups. It represents common knowledge shared via media such as textbooks and journals. Implicit and explicit knowledge are mutually exclusive, by definition, because once implicit knowledge becomes expressible it is no longer implicit but is rather explicit.

The second important distinction comes from AI, in which foreground knowledge ($F_k$) is differentiated from background knowledge ($B_k$) (Flach & Kakas, 2000). Foreground knowledge refers here to knowledge that is being newly developed in a situation—e.g. during a field campaign, laboratory study or other research activity—and consists of new situational knowledge as well as new implicit knowledge such as intuitions. Foreground knowledge might be viewed as an explanation being formulated by a scientist. In contrast, background knowledge refers to prior knowledge in the domain, including situational and theoretical knowledge as well as prior implicit knowledge. The distinguishing attributes are time and acceptance: background knowledge is knowledge that has been previously accepted as valid, whereas foreground knowledge is being actively promoted and tested. The relationships proposed here between these knowledge categories are stated in Eq. 2.1, Eq. 2.2. and Eq. 2.3, and are illustrated in Figure 2.5.

\[
\text{Eq. 2.1:} \quad F_k = S_k + I_k \\
\text{Eq. 2.2:} \quad B_k = E_k + I_k \\
\text{Eq. 2.3:} \quad E_k = S_k + T_k
\]
The socio-cognitive perspective on geoscience ontology then not only reinforces the need for socio-cognitively informed methods of knowledge capture, but also the need for multi-representations such as perspectives and versions, to represent varied explicit knowledge as well as implicit personal knowledge that might be inferred from contexts.

2.4.4 Scope of geoscience ontology

Of particular importance is the proposal here that the realist and constructivist approaches to knowledge overlap in the area of scientifically accredited explicit background knowledge (as shown in Figure 2.5). This is because the realist perspective ignores mental knowledge including its inexpressible parts (i.e. implicit), as well as the elements immediately participating in some knowledge discovery process (i.e. foreground knowledge), as it is more directed towards established scientific results that have a higher chance of being ‘TRUE’.

The scope of geoscientific ontology (or “geo-ontology”) can now be stated as consisting of this overlap area. In general this scope is intended to capture foundational geological knowledge that can subsequently be transformed to suit diverse scientific and societal purposes. The idea being that if foundational geoscientific knowledge, in all its partial and plural nature, is well represented it will not only stimulate new geoscientific research but also ease transformation to wider contexts within and beyond geoscience.

This scope intrinsically includes concepts, individuals and their states. Though all can be stored in databases, as data or metadata, it is much more common for individuals and states to exist in databases and for concepts to be expressed externally, e.g. in an ontology. The structure of such ontologies can then be refined to include concepts as well as semantic relations to corresponding individuals, to avoid duplication of content between the ontology and the database.

There are two semantic relations to be considered: extensional relations between a concept and its instances, and an intensional relation between a concept and its formative data. Extensional development involves linking data to the concept they exemplify and

---

**Figure 2.5: types of geoscientific knowledge.**

The overlap between realist and cognitive forms of knowledge.
corresponds to the database-mixed ontology mapping (dashed lines, right, Figure 2.4 above). Intensional development refers to the process and individuals involved in generating a concept—e.g. it can correspond to the mapping of a mixed ontology to the supporting, possibly prototypical, individuals in databases (solid line, right, Figure 2.4 above). In terms of relevance to ontology structure, inclusion of the extensional relations is optional because they mainly inform individual characteristics and existence, but inclusion of the intensional relations is mandatory because they amplify concept meaning by informing concept genesis, and because they have scientific merit by recording the data and reasoning processes used to derive a concept, which facilitates scientific re-examination and testing.

Scoping geo-ontology as per explicit knowledge suggests ontology development might proceed bottom-up, beginning with a catalog of explicit situational knowledge \((S_k)\) such as standard reference models and links to supporting premises. Alternatively, it might proceed top-down, with a catalog of explicit theoretical knowledge \((T_k)\) that transcends situations and is generally shared within geoscience, e.g. standard concepts, such as a normal fault, as well as widely accepted theories, e.g. plate tectonics. Experience indicates that any one approach in isolation soon ends up needing the other in practical applications.

Thus the following geoscientific constraint can be added to ontology challenges:

**Geo-ontology challenge 1:** capture, representation and use of explicit background knowledge consisting of situational elements such as reference models and prototypical individuals, as well as theoretical elements, such as theories and concepts.

### 2.4.5 Reasoning with geoscience knowledge

Understanding geoscientific knowledge that evolves and internally competes entails comprehending the factors leading to change or variety. These factors include not only the supporting evidence but also the reasoning process applied to it. Common forms of scientific reasoning can be framed using the knowledge distinctions above, i.e. foreground knowledge is derived from background knowledge and ancillary foreground knowledge (Eq. 2.4). The use of these distinctions allows scientific reasoning mechanisms to include the implicit human aspects that are so vital some science activities such as those field-based. When the contribution of implicit knowledge is marginal, or non-existent (such as in machine inference), then foreground knowledge can be viewed simply as explicit statements brought into focus for some line of inquiry.

\[
F_k B_k \rightarrow F_k 
\]

The first mode of reasoning is associated with situational knowledge development and hence with the generation of explanations and predictions. These coincide, respectively, with abductive and deductive forms of inference \((\text{Flach & Kakas, 2000})\). Due to the historical and causal nature of geoscience one of the \(F_k\) in Eq. 2.4 typically refers to a cause and the other to an effect. Reasoning forward from cause to effect (i.e. \(\text{cause } B_k \rightarrow \text{effect}\)), using background knowledge, represents deduction and prediction, whereas reasoning backward from effect and background knowledge to cause (i.e. \(\text{effect } B_k \rightarrow \text{cause}\)) represents abduction and explanation (see \(\text{Engelhardt & Zimmermann, 1982}\)). The common geoscience practice of developing models and process histories from observations is thus inherently abductive and
explanatory; using such models to develop future situations is deductive and predictive. From a logical viewpoint, abductive explanations are less certain because effects have the potential to be satisfied by multiple causes (i.e. competing models, hypotheses, concepts, individuals). The impact of this rule of logic is magnified in geoscience in that causes are often historical and unobservable.

Deduction on the other hand generates logical consequences when theory is applied to a situation; it further possesses the falsification property whereby false consequences imply false premises (i.e. \( \neg \text{cause} \rightarrow \neg (\text{effect } B_k) \)). Deduction is therefore superior to abduction from a logical viewpoint, but its certainty is also questioned when, from a pragmatic viewpoint, the theory and situation are often uncertain, incomplete, or indeterminate, allowing for variable predictive success.

The omission of \( B_k \) in Eq. 2.4 (i.e. \( F_k \rightarrow F_k \) or (cause, effect) \( \rightarrow F_k \)) represents generalization and induction, where empirical regularities are derived for the situation, e.g. the concept of ‘Baker Brook basalt’ as derived from field observations. Geoscientific generalization may be uncertain because, logically, inductions may be overturned with future contradictory evidence, and because, geologically, the underlying knowledge being generalized or abstracted may itself be uncertain, incomplete, singular, etc. The latter requires background knowledge, or the return of \( B_k \) to the logical form, to guide and validate the induction, leading to the claim that logical induction can be considered a weak form of abduction (Josephson & Josephson, 1994). This resonates with Peirce’s later pragmatic account of inference (1878) which is not based on logical form but rather on scientific activities whereby abduction generally refers to the discovery of new knowledge, deduction to prediction generation, and induction to the process of evaluating (dis/confirming) predictions within a broader population (Sowa, 2000).

Geo-ontology challenge 2: capture, representation and use of the geoscientific reasoning process behind some knowledge fragment, as part of its context.

2.4.6 Operating with geoscience knowledge

Scientific method has been studied by philosophers (e.g. Hempel & Oppenheim, 1988; Peirce, 1878; Popper, 1959) as well as geoscientists (e.g. Albritton, 1963; Harvey, 1969; Kitts, 1977; Martin, 1998; Schumm, 1991). Some of the results are empirically supported by psychologists (Feist and Gorman, 1998; Zimmerman, 2000) and implemented computationally (Langley, 2000; Shrager & Langley, 1990; Thagard, 1988), but with limited application to geoscience (e.g. Thagard & Nowak, 1990). The premise behind these computational efforts suggests that studies of scientific method may provide a template for the computer-aided discovery and use of geoscientific knowledge (e.g. Gahegan & Brodaric, 2002)—in essence, that they may provide a geoscientific metaphor for using information and tools in cyberinfrastructure.

Initially, philosophic approaches to scientific method were geared to the development of theoretical knowledge but subsequent critiques resulted in better accommodation for aspects of situational and implicit knowledge. The initial sequence can be summarized as consisting of: (1) neutral observation, (2) induction of problem statement or hypothesis, (3) deducing test criteria, (4) inductive confirmation of the hypothesis via further observation, and (5) communicating results (Hempel & Oppenheim, 1988). The sequence was amended by many including Popper (1959) who sought to mitigate its logical frailties by eliminating: (1),
because of value-laden observation, relegating (2) to psychological processes and advocating theory falsification rather than confirmation in (3), because induction is susceptible to future contradiction. Subsequent responses emphasized the validity of abduction as an inference form and the role of implicit as well as background knowledge in observation and reasoning (Feyerband, 1975; Hanson, 1958; Kuhn, 1962). This re-introduces (1) and (2), recast in abductive terms, and leads to the notion that theories might be valid if their parts cohere overall in spite of some internal uncertainties, incompleteness and conflicts. Geoscience method is often discussed in this recast framework (e.g. Schumm, 1991). The coherence principle implies potential non-monotonicity, i.e. where new observations could trigger modification or invalidation of existing knowledge. This is relevant to situational knowledge, as it is particularly susceptible to reinterpretation. Also significant is a knowledge-driven view of presentation, in which knowledge is transformed to suit a particular medium for communication purposes, e.g. a map as a visual presentation of explicit knowledge.

**Geo-ontology challenge 3:** use of metaphors and tools grounded in geoscientific method and use of knowledge-rich presentation mechanisms that link ontologies with, e.g. maps.

### 2.5 Ontologies and pragmatics

Pragmatics is proposed here as the area of knowledge representation generally concerned with how context influences meaning. A computational implementation of pragmatics then necessarily addresses at least ontology challenge 1, and geo-ontology challenges 1-2, all of which are concerned with context. In this section, a notion of pragmatics that applies to geoscience concepts (i.e. geo-pragmatics) is sketched as an important ingredient to context. This notion is later explored theoretically (chapter 3), empirically (chapter 4), and computationally (chapters 5, 6) in subsequent chapters.

#### 2.5.1 What is pragmatics?

Pragmatics is recognized as a branch in philosophy and in linguistics. In linguistic traditions pragmatics deals with the contexts surrounding formal and natural language use. For example, in natural language it refers to the impact on meaning of linguistic elements such as index terms whose reference can shift with situation (“I”, “you”, “it”) or to the meanings implied by tone or emphasis during actual utterance (Horn & Ward, 2004).

This focus on the contextual and usage aspects of meaning is also a cornerstone in pragmatic philosophy, where philosophers such as Peirce, Schiller, James, Dewey, Morris, and others, share the notion that significant issues can be best understood by studying their practical consequences. This particularly holds for C.S.Peirce, for whom the meaning of an idea requires three dominant threads: one for the experiential situation giving rise to an idea, a second for the definition of the idea, and a third for the actual and possible consequences entailed by the idea (1997/1878). Thus the pragmatic meaning of the concept ‘Baker Brook basalt’ (from Figure 1.1) can be supported by (1) field observations, (2) a defining description in terms of attributes such as age and rock type, e.g. “‘Baker Brook Basalt’ is extrusive and Ordovician”, and (3) the fact that from this description and other geological principles we can predict that it will likely overlay older rock bodies, and observe that it indeed does so. These threads are inherently tied to a scientific process in which they are discovered and used, making an idea a scientific artefact.
For Peirce such an idea is also anchored in a sign, which is a triadic relation between some object detected in the world, some representation of it (e.g. a symbol), and some abstraction of it by some interpreter (e.g. a concept). Figure 2.6 illustrates a sign in which the text “Baker Brook basalt” as written in a field notebook is related to some specific observed rock body and to the ‘Baker Brook basalt’ concept obtained in the mind of a geologist.

![Image of a sign for the concept 'Baker Brook basalt'.](image)

Any of the three nodes of the sign could be considered a departure point for the stimulation of a new sign in a process called semiosis. For example, the concept of the ‘Baker Brook basalt’ might serve in the role of object and stimulate another concept such as ‘formation’, which is represented in writing as “formation”. Or the object Baker Brook basalt might serve in the role of symbol, suggesting to a geologist both a concept of, and object for, some volcanic process operating in the region. Scientific inquiry is then the process of semiosis in which scientific reasoning guides the development of signs and converges them to a suite of signs that are scientifically credible and reflective of reality within the limits of such a lens.

Peirce’s three threads of a concept can now be interpreted in terms of signs and semiosis: experiential situations map onto the signs from which a concept is developed via semiosis; defining descriptions map onto the sign for which the concept is an abstraction; and consequences map onto signs derived from the concept via semiosis. Indeed, Morris mirrors this interpretation when he describes pragmatics as the “origins, uses and effects of signs” relative to an interpreter (1938). However, he differs from Peirce in two ways relevant to this dissertation: firstly, users and objects are not valid signs for Morris whereas for Peirce signs can be anything that provoke a response in an interpreter, including material objects, abstractions, representations, and users; and secondly, use seems to have superceded or encapsulated definition as a pragmatic dimension for Morris. In this dissertation the original
Peircean notion of signs is adopted because of its broader applicability to science, while the latter use-definition distinction is elaborated for knowledge representation purposes.

In particular, later developments in computational semiotics further advance the distinction between use and definition, assigning to semantics those aspects that are definitional, including potential uses, among other things, and to pragmatics those that are use-oriented, including actual uses, among other things (Stamper, 1996; 2000). Cartographic semiotics also reflects these distinctions, inasmuch as the meaning of a map, or its parts, involves the pragmatic notions of how the map originated, what it affected and how it was used, as well as the semantic focus on the relation between the map and the objects it denotes (MacEachren, 1995), including the possibly definitional characterization of the mapped objects.

So, adapting these distinctions to the ‘Baker Brook basalt’ results in semantics containing descriptions for typical characteristics, the possible functions it could enact or partake in, the situation types in which it could validly occur, and the relation to its instances (e.g. to the areas denoted by the polygons in Figure 1.1). In contrast, the pragmatics layer would then encompass those activities in which the ‘Baker Brook basalt’ is used (e.g. for environmental decision-making), the actual situations contributing to how it is used and discovered (e.g. field observations), and what actual effects result from usage (e.g. the designation of a protected wildlife area). Table 2.2 further adapts these notions to concepts for this dissertation, designating semantics as intensional and ahistorical, possessing typical characteristics, functions, constraining situations and a relation to instances, and pragmatics as extensional and historical, describing what functions are actually enacted, what situations are actually encountered to provide context, and what effects actually result.

Table 2.2: semantic and pragmatic aspects of concepts.

<table>
<thead>
<tr>
<th></th>
<th>origin</th>
<th>use</th>
<th>effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantics</td>
<td>situational:</td>
<td>definition: typical</td>
<td>potential results from use</td>
</tr>
<tr>
<td>(intensional /</td>
<td>operational</td>
<td>attributes, potential</td>
<td></td>
</tr>
<tr>
<td>ahistorical)</td>
<td>constraints</td>
<td>functions, instance</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>relation</td>
<td></td>
</tr>
<tr>
<td>Pragmatics</td>
<td>actual situations in</td>
<td>direct use of concepts</td>
<td>actual results from concept use</td>
</tr>
<tr>
<td>(extensional /</td>
<td>concept discovery</td>
<td></td>
<td></td>
</tr>
<tr>
<td>historical)</td>
<td>and use</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Another important aspect of computational semiotics is the role of the agents, who are seen as producers or consumers of signs, with the latter typically emphasized. Indeed Morris describes pragmatics as the relation of a sign to its consumer (or interpreter), which emphasizes the effects of signs on some recipient, but backgrounds the importance of the agents producing and using signs. Highlighting these additional roles and identifying who qualifies as a valid agent then impacts the pragmatics of concepts.

Firstly, designating humans as valid agents enables the pragmatics of concepts to include the human contributions to their origins, uses and effects (e.g. Stamper 1996; 2000). This would include, for example, encoding of the inferences and methods used within a scientific reasoning process, e.g. to discover a concept, as well as other human actions, related
intentions, communications, beliefs, etc. Broadening the scope to include machine agents allows pragmatics to consider how machine operations contribute to the pragmatics of concepts (e.g. Goldkuhl & Agerfalk, 2002), an increasingly important task given science’s reliance on computational tools for analysis, visualization, synthesis, etc.

Designating machine agents as producers of signs allows those signs to be interpreted by humans in a semiotic framework. Pragmatics thus retains a largely human bent, but one that lies at the heart of cyberinfrastructure as it implies that humans may interpret nature indirectly from its machine representation. But, if machine agents are also designated as consumers of signs, as seems quite reasonable, then machines can be seen as having an independent pragmatics in which machines generate and interpret signs. Indeed, the whole quest for automated interoperability of information, e.g. as in the Semantic Web, is predicated on this notion that machines can be quasi-intelligent producers and consumers of signs; many theories of human-computer interfaces are also predicated on this notion. Now, if we extend these ideas to consider scientific scenarios in which machines are automatically responding to input from sensors in nature, then machine pragmatics does not simply concern the interpretation of messages between machines, but also the interpretation of messages between nature and machines, analogous to how humans interpret natural signs. Therefore, both human and machine agents can have a dual nature, based on their role in the production and consumption of signs: humans can directly interpret natural signs, and indirectly interpret machine signs about nature, or even other human signs about nature (i.e. in language); likewise machines can directly interpret natural signs and indirectly interpret machine and human signs about nature. Of course, machine interpretation is much more limited than human interpretation, and indeed is constrained by the boundaries put in place by its human engineers, nonetheless within these boundaries machine interpretation can be seen as making choices hence interpreting signs rather than mechanistically reacting.

Finally, including natural agents (i.e. non-human natural entities) recognizes that nature itself can be a producer and consumer of natural signs, quite independent of human involvement. For example, biosemiotics is a discipline in which a wide variety of biological entities, of varying sentience, are seen as capable of producing, using and being affected by biological signs (Kull, 1999). It is questionable, however, whether non-biological entities can be semiotic agents. The issue can be expressed in terms of the role of agents in causality: in semiotic systems cause and effect are mediated by an agent who can variously interpret inputs, resulting in a range of potential actions and outputs (Emmeche, 2002). Recast in above terms, this expresses the idea that an agent is responsible for selecting a specific use and applying it within an origin for an indeterminate effect. Semiotic agents thus have some autonomy, or intentionality, in responding to stimulus and effecting results. Non-living systems are often described as not fitting this mould, because the relation between cause and effect is direct, determinate, and unintentional—i.e. not mediated by an agent, but fixed by a physical law. But, this may be a Newtonian assumption that may not hold at quantum or Einsteiniian scales, where the relation between cause and effect may be complex, suggesting the semiotic status of non-living systems is uncertain and possibly scale-dependent. What does seem to hold across all natural systems are the pragmatic aspects of origin, use and effect, and what seems contentious is whether uses are deterministically entailed by origins, and whether effects are entailed by both of these under some natural law.
At first glance natural pragmatics seems to be an interesting theoretical idea without any direct relevance to concept representation. It also does not seem very pragmatic at a practical level: if we marginalize the notion that geoscientific entities act as quasi-intelligent interpreters, because this view is seldom used in science today, then we are left with the sparse idea that geoscientific entities in nature can partake in cause-effect relations, i.e. in processes. If so, natural processes have no place in pragmatic representations, as they are a subject of pragmatic explanation and not its content. Is pragmatics then solely concerned with how humans and machines obtain and interpret knowledge? Without the recognition of natural situated concepts, the answer would likely be affirmative.

As discussed in more detail below, a naturally situated concept is one that reflects a pattern in nature that is local to a geospace-time region, hence is not universally applicable, and which is caused by local historical interactions of entities. The concept of a particular species, such as ‘human’, is one such concept (Millikan, 2000) and it is argued here that so is the ‘Baker Brook basalt’. The key point is that if the ‘Baker Brook basalt’ is indeed such a concept—i.e. a natural pattern of rocks found in multiple places within a specific geographic region, time interval, and with a common history—then its natural origins are central to its meaning, as independent of the scientific process of discovery as is possible. These origins differ from the origins of the various instances of the ‘Baker Brook basalt’, which would comprise its data and form part of a model, in that the concept’s origins describe what is common or prototypical in all instances, rather than what additionally unique factors were also involved in the creation of any one instance. The biological analogy would compare the origin of ‘humans’ to the origin of any one human. In the same way, common or prototypical uses and effects can be discussed for naturally situated concepts. Therefore a pragmatics for naturally situated concepts must include at least their natural origins (i.e. their natural pragmatics) and these origins correspond to, but also can differ from, histories of instances. Situated concepts thus enable us to distinguish between the natural origins for some concepts and the natural origins for individuals. This in turn might then allow us to differentiate two types of process models, one composed of individuals and their origins and one composed of situated concepts and their origins (see chapter 6), and each can be viewed as a pragmatic origin at their respective abstraction level. Entities in both process models can in addition have human and possibly associated machine pragmatics, to explain their discovery, etc. A situated concept such as ‘Baker Brook basalt’ or ‘human’ then has a natural pragmatics that involves a shared history amongst its instances, as well as a human pragmatics that involves scientific inferences, motivations, etc., associated with its human discovery, and perhaps a machine pragmatics in which instrumentation aids human discovery.

There exist some other obvious difficulties with a combined human-machine-nature view of agency in computational semiotics: i.e. machine and natural pragmatics lack the socio-cognitive aspects of human pragmatics, such as inherent beliefs, expectations, etc., which implicitly seem to frame interpretation, but nonetheless seem to act as semiotic hence pragmatic agents. Also, the three domains of human, machine and nature are somewhat interdependent, inasmuch as our knowledge of natural history is dependent on sampling and other human and machine-instrument constraints to scientific method, which could lead to history revision with additional data or theory. However, these limitations may act to parameterize a three-prong approach to pragmatics, rather than nullify them. Indeed, this dissertation takes this viewpoint as it argues that a complete notion of pragmatics for some geoscience concepts, i.e. naturally situated concepts, must involve all three types of agents.
The pragmatics of such concepts would then include the natural as well as the human and computational origins, uses and effects. For example, the pragmatics for the ‘Baker Brook basalt’ could include the human actions involved in discovering and using it, the machine operations deployed, and its process history including the key natural antecedents and descendants that share relevance amongst its instances. This combined approach is explored further in ch 3. Because of the dependence on geographical factors such as originating geographical area and environment, such pragmatics are called “geo-pragmatics” here.

2.5.2 Impact of pragmatics on semantics

The term “semantics” is used widely and diversely. So far it has been used according to one convention in information science whereby it refers to the layer of information associated with ontologies. Included in this layer is the relation between a concept and some represented entity, e.g. the semantics of some geospatial information in a database might refer to a concept in an ontology. The impact of pragmatics, as described above, on such semantics then involves enhancing ontologic content with pragmatic aspects in order to contextualize them. For instance, concept representations could be extended to accommodate their origins, effects and uses, in the human, machine and natural domains.

However, “semantics” is also commonly used to refer to the area of study of the relation between language and some entity in the world, e.g. “Baker Brook basalt is extrusive” and the observation that it actually is so in the world (e.g. Tarski, 1944). Several models exist for the semantic relation. Their significance here has to do with the interpretation of a concept within an ontology, and hence with the overarching principles and protocols that guide ontology design and use, including the potential influence of pragmatics.

For example, in realist semantics there is a direct relation between the world and abstractions, and between mind and world, as shown in Figure 2.7 (Grenon & Smith, 2004; Smith, 2004). An abstraction is present in, and given meaning by, each world instance to which it applies, and the relations between abstractions reflect relations in the world (Bittner & Smith, 2003). Importantly, abstractions and the world are directly assessable for representation in humans, machines, etc., and ontologies are then digital representations of abstractions. Apart from the issue of how abstract entities, such as numbers, fit into this scheme, this approach leaves limited scope for pragmatics to influence ontologies: only natural pragmatics might apply, because world-abstraction relation is primitive, and to some degree human pragmatics might apply, because representation is somewhat recognized in the semantic relation. The latter allows for annotation of an abstraction with metadata about its representation, e.g. encoding, use, etc., but not necessarily about its discovery. Also, only minor reference is given to discrepancy in representation, such as diverging scientific opinions, which are seen as partial or in error. In essence, there is no theoretical basis to describe how abstractions are discovered via scientific process because the theory is mainly concerned with results of abstraction and not its process (Fonseca & Martin, 2005). Including the scientific process of discovery in ontologies is then theoretically antithetical. A variation of this program occurs in the semantics of natural or artificial languages, which primarily consider the relation between language and the real or some possible world.
In cognitive semantics, abstractions and language are both given meaning in the mind, and semantics refers to the relation between mind and world (Figure 2.8; Gardenfors, 2000). Concepts in ontologies are thus digital representations of mental entities, and pragmatics might involve representations of the mental entities that contextualize concepts. Morris’ semiotic formulation of semantics is similar, in that abstractions are produced and housed by interpreters, usually humans, and semantics refers to the relation between world and sign. Because of the mental orientation both obviously allow for human pragmatics, but are necessarily challenged by natural pragmatics and scientific pragmatics due to variable access to the world from mind. Consequently, in this framework it is difficult to theoretically separate scientific result from opinion.

To overcome the limitations of cognitive and realist semantics, some authors (e.g. Millikan, 2000) have proposed a hybrid cognitive-realist semantics in which some abstractions can be anchored in the relation between world and mind, as shown in Figure 2.9. This approach in general provides access to world from mind, at least for those abstractions anchored by natural necessity, and this allows for scientific knowledge of the world, as well as for natural, human and scientific pragmatics, to be theoretically incorporated into ontologies. It is also compatible with Peircean inquiry in general, in terms of enabling access to world via science, though the connections between mind and world are discussed in different ways. The view is
also compatible with critiques of the realist approach that argue for the need to incorporate the scientific process into ontology design (e.g. Fonseca & Martin, 2005).

Using Millikan’s language, a concept is then anchored in reality for substances such as objects and materials, a kind is the entity in nature that anchors the concept, and a conception is a mental entity. For this dissertation, this notion of concept is extended to the realm of science such that a scientific concept is one obtained via scientific process. This makes intuitive sense: not all kinds are recognized by science, but all concepts do apply to kinds, and certainly multiple concepts can apply to a kind, perhaps from different perspectives. Furthermore, not all conceptions are necessarily concepts, not all concepts are picked up by everyone, and some things that are picked up are not necessarily evaluated to be concepts—thus, one scientist can detect a pattern with valid correlates in nature but disregard it, while another can value it. What is unclear is the status of the disregarded entity—is it concept, conception, or neither? In this dissertation it is considered to be a conception until scientifically accredited as a concept, mainly because there seems to be no clear alternative.

Lastly, Millikan suggests that language is a valid interface to concepts, on par with other sensory inputs. If we extend this to include representations in general, then this line of thought evolves to the semiotic argument that signs, and sign interactions through semiosis, can reveal concepts, which altogether aligns with Peirce’s general thrust.

Consequently, in this dissertation, semantics is aligned with the precepts of cognitive-realist semantics, but follows the information science convention in which semantics refers to the information layer that contains ontologies. Also, following Peircean pragmatics, abstractions are considered to be scientific artefacts and are referred to as concepts.

Concepts are then abstractions of varying generality, such as ‘Baker Brook basalt’, ‘geological formation’, ‘granite’ or even ‘process’. In particular, geospatial concepts refer to abstractions about entities associated with geographic space, such as types of geospatial relations (e.g. near, on, beside,…), objects (e.g. points, lines,…), properties (e.g. curvilinear, open, closed,…), positions (e.g. coordinates, spatial reference systems…), while geoscientific concepts refer to abstractions about natural entities occurring in geographic space and discovered via scientific investigation, e.g. geologic units, ecologic units, hydrologic units, soil units, their associated properties and processes, etc. Here, the occurring relation broadly refers to entities located directly in geographic space, such as physical objects, or indirectly
via relations to objects such as those held by an object’s properties or the processes the object partakes in (Grenon & Smith, 2004; Gangemi, et al., 2003).

Geoscientific concepts are also considered here to be part of the broader category of geographical concepts, which include abstractions about both natural and non-natural entities occurring in geographic space. Natural and non-natural concepts are distinguished here by the primary origins of the entities for which they are abstractions: natural entities are created by nature, whereas non-natural entities are dependent on humans or machines for their creation. Significantly, geographical and geospatial concepts are related in the following way: geographical objects possess properties that refer to geospatial entities (Gangemi, et al., 2003)—e.g. geoscientific mapped units have geometry, which refers to a geospatial object, and have location which refers to a geospatial position. Representations of the geographic domain consequently need to include both geographical and geospatial concepts.

In summary, in this dissertation, concepts refer to abstractions that are scientific artefacts operating under cognitive-realistic semantic assumptions, and three types of concepts are distinguished: geospatial concepts refer to the abstractions about geographical space, geographical concepts refer to abstractions about entities occurring in such space, and geoscientific concepts refer to geographical concepts that have natural origins. Conceptions refer to mental entities, and geo-pragmatics refers to the pragmatic aspects of geoscientific artefacts.

Theories, on the other hand, are coherent (non-contradictory) collections of expressions that help define concepts, such as ‘‘Baker Brook basalt’’ is extrusive and Ordovician” or “if ‘Baker Brook Basalt’ then….” When concepts and related expressions are coherently organized into a collection, the collection is referred to as an ontology. An ontology can then be viewed as a specific theory about its constituent concepts. As noted above, it is important to distinguish between philosophical ontology, which refers to a prior concepts that are foundational and ‘true’, and computational ontology which refers to an engineering artefact consisting of digitally represented concepts usually encoded in some logical formalism (Smith, 1998a,b). Reference to ontology in this dissertation is aligned with the latter, i.e. an engineering artefact, but one guided by the former, namely the general principles of pragmatics, which includes the cognitive-realistic assumptions. Following these principles, an ontology is then a type of scientific theory, one that is formally represented and that explicates certain scientific concepts recognized by humans. This view inherits certain representational benefits, due to its engineering viewpoint, that enables both abstractions and the world to be symbolically represented. Abstractions might then be represented using RDF-OWL, and world objects might be held in a database and accessed using a database language such as SQL: e.g. the ‘Baker Brook basalt’ definition can be expressed in the RDF-OWL syntax shown in chapter 1, and this snippet can be related to its instances residing in a geospatial database (e.g. to polygons 102 and 790 in Figure 1.1) or in some other RDF-OWL document.

Turning to the role of digital world in such structures, it is apparent that world in this sense could refer to any cyber-resident entity over which abstractions can be made within a course of inquiry, that is, any entity about which concepts can be formed and tested during cyber-based scientific investigation. This approach leaves open the possibility that concepts can be a source of abstraction, that is, serve in the role of world in the semantic relation, as
suggested by semiotics. As shown in later chapters (3,5,6), a pragmatic view of concepts actually forces this issue, suggesting a dual role for concepts as both abstractions and world.

This dual role is studied in this dissertation primarily from an ontology design perspective, where it is argued that a dual view results in a clearer representation of the geoscientific domain. The main substantiation for this is the identification of a type of concept, the situated concept, that sits on the border between abstraction and world, and could operate both ways in an ontology. Its primary characteristic is that it applies only within a specific geospace-time region, and that its meaning is tied to the historical-causal connections of its members (after Millikan, 2000). Situated concepts are not applicable universally but only locally. They are indelibly tied to the pragmatic viewpoint, because their meaning is critically dependent on local interactions in a geospace-time circumstance, i.e. on their actual origins, which essentially means their histories, including their input state, geospace-time locale and environment, and production process. In other words, a necessary and sufficient condition for situated concepts is origin-dependence: if a concept is origin-dependent then it is situated, and if a concept is situated then it is origin-dependent.

Examples of situated concepts include specific species concepts (e.g. ‘human’) as well as specific geographical regions such as ecologic, geologic (e.g. ‘the Baker Brook basalt’), and others that are origin-dependent (for the origin-dependence of ecologic concepts see Fonseca & Martin, 2004). Apart from these examples taken from the purely natural domain, situated concepts also include concepts taken from the social domain. The key element to social-driven concepts is the involvement of human agency in their origins: for example, Millikan (2000) suggests that artefacts such as ‘Toyota Corolla’ are situated (or “historical”) because their production is geographically localized and tied to a specific corporate legacy. Shifting this to the geographical domain, a social-driven situated concept might be any limited generalization made by a human geographer about a case study in which the local conditions are explored from the perspective of unique defining histories—indeed, one might say that the purpose of such studies is to identify such concepts. A different example of a social-driven situated concept is ‘polluted lake’, which is tied to local regulatory conditions, shifting measurement protocols and instruments, and expectations about what the local environment can tolerate. The main point being that although the situated concepts considered herein will be mostly from the natural sciences, they indeed cover any concept that is dependent on its actual historical origins, including some social concepts.

Figure 1.1 contains a good example of a situated geoscience concept in the ‘Baker Brook basalt’: it is found within a certain geographic and temporal region as a result of its natural history, has multiple instances (see polygons 102 and 790 in Figure 1.1), but is not universally applicable like the ‘formation’ concept, which might apply to any region in geospace-time. Yet, the ‘Baker Brook basalt’ is only provisionally a concept: it seems to be an abstraction over multiple instances, where each instance could represent a geographically distinct rock body derived from a common generative event and original body of material. However, it is not clear in Figure 1.1 whether each polygon instance is indeed physically separate from the others, and hence whether the ‘Baker Brook basalt’ is in fact an abstraction over several instances and should thus be considered a concept, or whether the polygon instances are actually physically connected underground, with the map simply illustrating outcroppings, and hence whether the ‘Baker Brook basalt’ should instead be seen as a single world entity and perhaps not a concept. What is at question is the status of the physical unity
of the various polygons labeled as ‘Baker Brook basalt’, and in a deeper sense, whether abstractions over single entities or multiple entities can be concepts?

The answer to the first part of this question, whether the ‘Baker Brook basalt’ is a single rock body, is unknowable without more data and knowing its 3D layout. However, the answer to the deeper question actually overrides this issue: following a scientific realist position on this issue, an universal is an abstraction that applies to one or more numerically single concrete entities in nature, usually individuated in geospace-time, which are called individuals or particulars (Armstrong, 1978). Furthermore, there is no reason why this position cannot hold under the cognitive-realist assumptions taken here. The key point is that abstractions, hence concepts, can apply to a single entity or to multiple entities, i.e. we can respectively abstract about the changes of state over time of a single rock body, or about the commonalities of several rock bodies at one timepoint or over many timepoints. For the purposes of this chapter, and chapters 2-5 (but not chapter 6), this implies at least three types of entities: concepts that apply to multiple natural entities, individual concepts which abstract over states of a single natural entity over time, and states of individuals which describe the characteristics of an individual at some time. So the ‘Baker Brook basalt’ is a concept, regardless of its unity conditions. What is unknown is whether the tabular description of its instances in Figure 1.1 refers to the concept or its states; to hazard a guess, it is likely the former given that each description is identical.

However, having established that the ‘Baker Brook basalt’ is a concept, it is also possible to see it as an instance of the more general concept ‘geological formation’, which can apply to any geospace-time region. The rationale for this argues that ‘Baker Brook basalt’ is not purely abstract, as it is geographically, genetically and historically localized, and that it is indeed an example of the abstract universal ‘formation’. In fact, situated concepts such as these are weak universals and weak world entities, occupying an intermediary position. As such, ‘Baker Brook basalt’ could play a dual role in a geoscientific ontology: from the viewpoint of any of its instances it could be considered a concept, but from the viewpoint of the universal ‘formation’ concept it could also be considered a source of abstraction.

Valid subject matter for pragmatic-driven semantics then includes the internal nature of concepts, and their external relations with each other and the world. A pragmatic view of concepts impacts this subject matter inasmuch as it suggests that concept structure should enable tracking of a concept’s actual origins, in addition to supporting the representation of actual uses and effects, in the human, machine and natural domains. Although all three pragmatics aspects are important, this dissertation focuses on representing the origin aspect primarily, and the effect aspect secondarily. This emphasis is chosen because of the significance of these aspects to situated concepts, and to sharpen the focus of the dissertation. Issues with representing the use aspect of pragmatics are therefore considered mainly out of scope. In terms of pragmatic impacts on semantics, a geo-pragmatic viewpoint also implies a geographical ontology organization in which situated concepts occupy a position sandwiched between individual world entities and more general concepts that are not origin-dependent. Finally, the semantic relation itself is affected by a pragmatic view of concepts. Such a view draws attention to the need to maintain a link to the world entities involved in the origin of a concept, in addition to those entities that are instances of the concept, suggesting a complexity to this relation that is further explored in chapters (3,5).

**Geo-pragmatic challenge 1:** augmenting geoscience semantics with geo-pragmatic aspects.
2.5.3 Impact of geo-pragmatics on schematics

A pragmatic exploration of the meaning of geographical concepts impacts not only their semantics, but also their associated schematics, i.e. the structures used to represent concepts. Key geo-pragmatic issues that affect schematics include: the contribution of the geo-pragmatic aspects to the meaning of a naturally situated concept; the dynamic-historical nature of naturally situated concepts; and the dual role that a situated concept can play in any representation.

The first issue implies that any representation framework for concepts must make provision for at least the actual origin of a situated concept in its concept structure. The second issue implies that representation systems must allow situated concepts to be added and modified. Although this has some bearing on ontology systems, the overall impact on their schematics is minor because ontologies are basically intended to manage concepts. It does however have significant bearing on database systems, because databases are primarily intended to manage information, not concepts, and they assume that all concepts are predetermined and static. One approach to overcoming this in database systems is to treat at least situated concepts as a form of information to be recorded in the database, causing the database schema to be adjusted accordingly. These adjustments are explored in chapter 5, where it is suggested that a geospatial database must minimally store: (1) a characterization of the situated concept—i.e. often considered to be part of the scientific interpretation, (2) the information leading to it (its actual origins)—i.e. the observations, measurements, and prior knowledge, as well as (3) the instances of the concept—i.e. the actual regions that are often portrayed on a map. Although not considered herein, spatial data transfer standards would also need to be adjusted accordingly. The most recent proposal by the Open GIS Consortium (OGC; www.opengis.org) for such a standard, i.e. Geography Markup Language 3.0 (GML3), theoretically supports such adjustments, but technologies to deploy GML3 are almost non-existent, requiring geoscientific standards efforts to engineer various work-arounds in these technologies to preserve as many adjustments as possible (e.g. Cox, et al., 2005).

This treatment of concepts as a type of information in a database or data transfer standard also addresses the third issue in which situated concepts play a dual role in representations, as concepts and instances. The impact of this third issue on ontology schematics is marginal, but the impact on overall ontology structure is considerable and is discussed in chapter 6.

Geo-pragmatic challenge 2: augmenting geoscience schematics with geo-pragmatic aspects.

2.5.4 Impact of geo-pragmatics on syntax and sigmatics

The impact of pragmatics on syntax and sigmatics becomes most evident when they considered from a scientific viewpoint. Then, syntax is associated with “data”, which is understood to refer to the raw format and patterns, respectively, of the messages in which observations and measurements are expressed, and sigmatics is associated with “information”, which is understood to refer to a structured version of these inputs. Data are exemplified by the sequences of raw digital bits coming from a sensor, such as an imaging satellite or geophysical instrument, and information is exemplified by its storage in a specifically formatted digital file. Analogously, human perceptions during fieldwork might constitute data, which is turned into information via its lexical encoding in a field notebook. These definitions are often used to differentiate these notions in the knowledge management community (e.g. Tuomi, 1999).
Pragmatics could then affect syntax through: (1) operational changes that alter the format of the messages, such as changing the amplitude and intensity of the signals being received or the grammar of the languages in which messages are expressed, or (2) content changes in which patterns in the incoming messages shift with variances in natural situations, e.g. changes in the patterns of expression for the image-based or human-observed data of some landscape such as when moving from one geological unit to the next, from the pink ‘Buchans Group’ to the blue ‘Roberts Arm Group’ in Figure 1.1. The latter represents a shift in information content, hence in sigmatics, and investigating these affects might then provide empirical evidence for the impact of the origin aspect of geo-pragmatics. So, if methodology remains consistent and invariable during some scientific enterprise, thus nullifying (1), then inspecting scientific data for inherent patterns and shifts over time, and correlating these with the resultant situated concepts they support, should provide insight into the nature of actual origin dimension of those concepts. When the data are geographical observations and the concepts are geographical regions, then such inspection should inform on the actual origins of naturally situated concepts. This hypothesis is explored further in chapter 4.

**Geo-pragmatic challenge 3**: evaluating the presence of geo-pragmatic artefacts in data.

### 2.5.5 Geo-pragmatics versus metadata

The relation between metadata and geo-pragmatics is superficially one of overlap, but on deeper inspection the overlap is minor. Metadata as conceived by users and producers of geospatial products, and as defined by various standards bodies such as ISO/TC211 ([http://www.isotc211.org/](http://www.isotc211.org/)) and FGDC ([http://www.fgdc.gov/](http://www.fgdc.gov/)), typically includes information about geospatial information instances, called “features”, and their collections (i.e. about sigmatics), as well as information about data structure (i.e. about schematics). It includes aspects such as geospatial accuracy (+- 1 meter), scale (1:50,000), resolution (1 meter/pixel), completeness, consistency, lineage (version 2.1), projection (UTM), authorship, use constraints, and many others (Gunther & Voisard, 1998). Most notably, metadata is not yet commonly applied to concepts and indeed semantics is often vaguely described as a type of metadata that clarifies the meaning of data and schema by defining relevant concepts. The metadata element that does loosely correspond to the pragmatic aspect of origins is lineage, which is often called provenance. However, in spite of this single correspondence, the overall similarity of pragmatics to metadata is weak and there are important differences.

Firstly, geospatial metadata is typically assigned to individual geospatial features (e.g. polygon 102), or to collections of features (e.g. geospatial products such as maps), but not to the concepts used to classify features (e.g. ‘Baker Brook basalt’). Such metadata is therefore not oriented to explaining the meaning of features via additional explication of their underlying semantics or pragmatics, but is rather directed at explaining the pertinent operational lineage for information.

Secondly, a focus on operational feature lineage narrowly considers the human actions taken, or the human aspect of actual origins for features, but largely ignores the natural origins in theory, and the machine origins in practice, involved in producing features and concepts. It also provides no well-defined model for scientific inferences and operations involved in origins. This limits its use within the digital scientific process.

Thirdly, even when lineage, or provenance, is considered, it only represents the origins aspect of pragmatics and omits the effect and use aspects.
Fourthly, because of the human orientation and its lack of connection to natural, machine and scientific process, the majority of metadata representations are not rigorous, but are typically limited to free-form text descriptions or at best to keyword-based descriptions. Although there are exceptions to this generalization, and in spite of the fact that such descriptions might aid human understanding and are a vast improvement to previous practices that ignored metadata, by and large this overall situation does not greatly advance automated computation, such as automated search, nor does it greatly aid scientific processing inasmuch as it captures only a small fragment of the scientific process.

In contrast to metadata, geo-pragmatics as considered in this dissertation does apply to concepts as well as features, and does include the scientific, machine and natural process involved in discovering both geographical features and concepts. Because of this orientation, a geo-pragmatic representation can aid scientific computation in cyberinfrastructure inasmuch as scientific theories, including ontologies, could then be tested and evaluated against data, and new theories could be inferred from existing knowledge, via digital aids.

2.5.6 Geo-pragmatics and explanation

Geo-pragmatics as described here is closely connected to the notion of explanation. Explanation, including scientific explanation, strives to answer the question “why?” which encompasses notions of how, when, where, etc. Broadly speaking it answers such questions as “why is some conclusion reached”, or “why does some state or activity exist or persist”? In the sciences this largely involves a recounting of the origins or effects of the item being explained. For example, various law-driven or probability-driven models for scientific explanation explain a conclusion by recounting the premises—i.e. origins—in the inference leading to the conclusion. Alternatively, functional and rational-choice models explain something according to its actual or possible effects. Both these types of models have been criticized as being overly mechanistic and ignoring motivational, and other contextual factors: e.g. a rock body is represented with a certain shape because of: (1) its origins, including natural processes and materials as well as observation methodologies motivated by cost and availability, and (2) its effects, including its suitability for waste disposal resulting in excavation of some of its materials to host the waste. Rather than emphasizing any one of these types of explanation, geo-pragmatics embraces them all. It includes origins and effects, as well as machine, natural and human aspects, with the latter including motivational and related factors.

Geo-pragmatics as emphasized here also differs from explanation in its focus on concepts. Existing studies in explanation-based reasoning, and more recently their application to ontology, are somewhat addressing the origin dimension of concepts by tracking an inference, and inference chain, to explain some reasoning about concepts (e.g. McGuinness & Pinheiro da Silva, 2004). For example, if a conclusion is deductively drawn from a premise consisting of a condition plus some input state (e.g. the condition “if ‘Baker Brook basalt’…then…” and the input state “‘Baker Brook basalt’”) then the conclusion is explained by the premise. Moreover, this notion of premise and conclusion, hence explanation, can be generalized to reasoning mechanisms beyond those deductive and generally rule-based, where the condition of the premise is some procedure rather than a rule, e.g. reasoning from examples by determining similarities (e.g. case-based reasoning; Aamodt & Plaza, 1994) or a simulation or modeling process. By and large the explanations generated in all these efforts are for machine inferences, and can be dynamically created by reconstructing the reasoning.
chain, by “proving” the conclusion. This becomes problematic for explaining complex inferences, such as those generated by non-machine agents, which cannot be dynamically reconstructed because they are generated elsewhere (e.g. in humans), are too complex (e.g. diagrammatic reasoning), or are non-deterministic and cannot be reduced to a simple causal reasoning chain (e.g. social or neural networks). In these cases it would be useful to declaratively store an explanation/origin for some conclusion. A case in point is any scientific conclusion, inasmuch recording the origin would allow the conclusion to be validated, replicated and evaluated.

2.6 Summary of assumptions and challenges
A listing of the major assumptions made in this dissertation is given below. Also listed are the significant challenges identified in this chapter. The assumptions apply throughout the dissertation, but only some of the challenges are addressed, in part, by the dissertation. Connections to the objectives as set out in chapter 1 are also given below. Note that this chapter itself fulfills the first objective of characterizing some challenges to geo-interoperability.

2.6.1 Assumptions
The following major assumptions are made in this chapter and apply in this dissertation:

- **Semantics**: a cognitive-realist semantics is assumed in which some abstractions are anchored in the relation between world and mind. This assumption provides a theoretical-philosophical basis for scientific inquiry and related ontology development.

- **Concepts**: are scientific artefacts resulting from scientific inquiry. This entails that the discovery and evolution of a concept should factor into its meaning and be represented. *Concepts* are purely mental artefacts.

- **Pragmatic agents**: humans, machines and nature are valid pragmatic agents that can produce signs and variably consume signs. Natural agents consume only natural signs, whereas machine and human agents consume natural, machine and human signs.

- **Geo-pragmatics of concepts**: involves dimensions for origins, effects and uses, within the human, machine and natural domains for geoscientific concepts. This assumption provides a meta-structure for concept representations that heeds scientific discovery and evolution.

2.6.2 Challenges
The following challenges to geoscience knowledge representation are identified in this chapter. Each group of challenges refines some aspects of the group preceding it, narrowing the scope and focus progressively. Note that the objectives of this dissertation as described out in chapter 1 mesh directly with the last group of challenges concerning geo-pragmatics.

**General ontology challenges**

- **Ontology challenge 1**: Developing and using social, cognitive, lexical, and philosophical methods for capturing ontologies inherent in groups, heads, documents and the world.
This challenge highlights the variety of knowledge repositories and the need to develop and deploy unique methods to mine them. Failure to meet this challenge will result in ontologies with content deficits.

- **Ontology challenge 2: Representing perspectives, versions, contexts and graded concepts.**

  This challenge is central to this dissertation. It highlights the impoverished nature of current modes of representing concepts in ontologies and seeks to augment those primarily by introducing constructs that enable concepts to be contextualized for scientific applications. Failure to meet this challenge will likely result in ontology structures that do not fit how scientists, and others, structure knowledge.

- **Ontology challenge 3: Developing ontology-enabled workflows from scientific activities.**

  This challenge highlights the need to transfer scientific activities into cyber environments. Failure to do will result in cyber environments that are not useful because they cannot be understood or because they simply do not suffice. In many ways the future of cyberinfrastructures rests in rising to this challenge.

**Geo-ontology challenges**

- **Geo-ontology challenge 1: capture, representation and use of explicit background knowledge consisting of situational elements, such as reference models and prototypical individuals, as well as theoretical elements, such as theories and concepts.**

  This challenge highlights some of the unique characteristics of geoscience, and geography as a whole, in which reference observations and models are critical knowledge pieces. This leads to an acknowledgement of the significance of situated concepts, which are generalizations of some geospace-time situations, and represents a major contribution to the conceptualization of ontologies.

- **Geo-ontology challenge 2: capture, representation and use of the geoscientific reasoning process behind some knowledge fragment, as part of its context.**

  This challenge effectively calls for the incorporation of geo-pragmatic factors into concept representation in order to boost context. The critical factors involve tying concepts to modes of scientific reasoning, and exploring the unique characteristics of geoscience in which models are inferred much more often than theories.

- **Geo-ontology challenge 3: use of metaphors and tools grounded in geoscientific method and use of knowledge-rich presentation mechanisms that link geo-ontologies with, e.g. maps.**

  This challenge refines the second challenge above, in that it calls for use of mechanisms grounded in geoscience, which should involve map-based activities and geospatial simulations, etc., as well as the ability to connect them in general and custom workflows.

**Geo-pragmatics challenges**

The following challenges cumulatively refine both the first and second geo-ontology challenges from above, by exploring different aspects of enhanced context representation for geoscience concepts, based on their origins, effects and use.
• *Geo-pragmatic challenge 1: augmenting geoscience semantics with geo-pragmatics.*

This narrows the scope of context to include how concepts have been discovered, used and what they have affected. Because it is concerned with semantics, this challenge is aimed at the idea that there exists a type of concept that is dependent on geospace-time situations. This meshes mainly with objective 5 (chapter 6) and somewhat with objective 2 (chapter 3), which aim to explore the theoretical boundaries of the idea.

• *Geo-pragmatic challenge 2: augmenting geoscience schematics with geo-pragmatics.*

This concerns the structuring of enhanced concept representations based on geo-pragmatic elements, and meshes with objective 2 (chapter 3), which lays out the theory, and with objective 4 (chapter 5), which applies the theory to database schema.

• *Geo-pragmatic challenge 3: evaluating the presence of geo-pragmatic artefacts in data.*

This concerns obtaining and evaluating the empirical evidence for geo-pragmatics, to gauge fit of theory to data. In general, the geographical discovery process is an area under-studied in all its forms, though in fact this step is required to substantiate any theory of geo-pragmatics. This challenges meshes with objective 3 (chapter 4).

The remainder of this dissertation will now focus on addressing the three geo-pragmatic challenges, thereby also addressing some aspects of the challenges to geoscience ontologies and other ontologies.
CHAPTER 3: REPRESENTING GEO-PRAGMATICS

This chapter introduces the pragmatic aspects of geoscientific representation, i.e. geo-pragmatics, develops a formalism and applies it using logic to a realistic example from geoscience. In doing so it lays the theoretical groundwork that will be empirically tested and computationally applied in subsequent chapters. By outlining a formal conceptualization for geo-pragmatics it responds to the second objective described in chapter 1, and the first two geo-pragmatic challenges described in ch 2, which call for concept representations to be augmented with geo-pragmatic elements.

The chapter proceeds as follows: section 3.1 discusses traits of scientific pragmatics in more detail; section 3.2 elaborates these traits for geoscientific pragmatics, or “geo-pragmatics”, and develops theory for their application to concepts; section 3.3 describes the elements involved in representing concepts using the theory; section 3.4 presents a sketch of a logic formalism for representing the theory; section 3.5 presents a geological example that demonstrates the theory in narrative as well as in logic form; and section 3.6 concludes with a short summary statement.

3.1 Scientific Pragmatics

In chapter 2, geo-pragmatics was defined in terms of three aspects, the origins, effects and uses of geographical entities, and these were discussed in terms of three domains of application, natural, human and machine. For example, an augmented pragmatic representation of the origins of a geoscientific artefact could involve: a process history of the phenomena as conceived of occurring in nature, the human activities involved in discovering the natural process history, such as various scientific methods, as well as the instruments involved in the discovery process, such as computational operations and workflows, laboratory analyses, etc. This suggests that increasing the level of meaning for a digital scientific artefact in terms of its origin involves not only capture of the natural evolutionary history of a phenomenon, but also capture of the methods and tools used by humans to discover it.

Table 3.1 illustrates the origin, effect and use dimensions of pragmatics in the human, instrument and natural domains. Note that a scientific viewpoint is taken: actions in the human domain involve the reasoning mechanisms used to discover natural phenomena (epistemological), actions in nature involve phenomena viewed by scientists as natural processes (ontological), and instrument actions involve machine operations and workflows that support human reasoning and detection of nature (methodological). Thus, for example, given that a scientific inference can be syntactically structured as premise $\rightarrow$ conclusion, then the origin of an artefact refers to the premise for which the artefact is a conclusion, the effect refers to a conclusion in which the artefact is found in the premise, and use refers to any logical sentence involving the artefact that potentially could be utilized in an inference. Similarly, given that a computational operation can be structured as task (input, output), then the origin of the digital representation of an artefact refers to the task and input for which it is an output, the effect refers to the task and output for which it was an input, and use refers to any task for which it can be an input. Natural events are analogous to computational operations except they are often stated in causal terms, i.e. event (cause, effect), where the causes and effects designate the before and after states of a transition event. Both tasks and events can occur within broader processes that serve to
sequence tasks and events—these processes are not included in Table 3.1 as their causal impact on origins is indirect. However, sometimes the natural origin of entities cannot be attributed to a discrete event, but rather to a continuous process (Sowa, 2000, pp. 213-217), i.e. process (input, output), where outputs designate states occurring anytime within the lifetime of the process. In that case it is the process that would be used to for the natural origins of an entity.

Table 3.1: pragmatics for a scientific artefact X in the scientific, machine and natural domains.

<table>
<thead>
<tr>
<th></th>
<th>Origin of X</th>
<th>Effect of X</th>
<th>Use of X</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scientific inference</strong></td>
<td>in Y ⊨ X:</td>
<td>in X ⊨ Y:</td>
<td>any statement involving X:</td>
</tr>
<tr>
<td>premise</td>
<td>Origin(X) = { ⊨, Y }</td>
<td>Effect(X) = { ⊨, Y }</td>
<td>Use(X) = { X,Y,Y ⊨ X, X,Y ⊨ Y }</td>
</tr>
<tr>
<td>conclusion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Instrument operation</strong></td>
<td>in task1(Y,X):</td>
<td>in task2 (X,Y):</td>
<td>any task involving X:</td>
</tr>
<tr>
<td>task (input, output)</td>
<td>Origin(X) = { task1, Y }</td>
<td>Effect(X) = { task2, Y }</td>
<td>Use(X) = { task1, task2 }</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Natural process</strong></td>
<td>in event1 (Y,X):</td>
<td>in event2 (X,Y):</td>
<td>any event involving X:</td>
</tr>
<tr>
<td>event (cause, effect)</td>
<td>Origin(X) = { event1, Y }</td>
<td>Effect(X) = { event2, Y }</td>
<td>Use(X) = { event1, event2 }</td>
</tr>
</tbody>
</table>

The notion of use as just described is quite loose—e.g. for a concept, use refers to any statement in which the concept is deployed. In reasoning, use might then refer to an axiom, definition or inference involving a concept, whereas origin and effect might refer to the deployment of the concept specifically in the premise or conclusion of an inference, respectively. Likewise in computational systems use can refer to the overall function of the procedure, denoted by its name (e.g. task above), in which the concept is used as input (i.e. anything accessed by the procedure) or output (i.e. anything changed or created by the procedure), whereas its analog in nature might refer to the event or process operating over natural entities. Use can thus refer to the overall activity that frames premise and conclusion, cause and effect, or input and output, e.g. when we consider not how the ‘Baker Brook basalt’ was discovered, nor how its various instances are determined, but rather its role in ‘waste disposal’ or ‘groundwater management’.

An outstanding issue now arising in digital scientific pragmatics is the interleaving of the human, instrument (including computational) and natural domains. In a scientific machine-driven cyberinfrastructure these elements can be related for representation as follows:

- A scientific inference maps onto an instrument operation such that:
  - premises map onto inputs
  - conclusions map onto outputs
  - inference operators map onto operators for computation or observation
- Natural entities, including objects and processes, map onto instrument inputs and outputs, either in their material form (e.g. samples for lab analysis) or in their digital representation (e.g. database record).
Figure 3.1 shows how these mappings (right) apply to a cause-effect relation in nature between two objects ($O_1$, $O_2$). The mappings are designated as being syntactic, because they denote how natural and human signs are represented in machines. Also shown are pragmatic mappings (left, middle) within and between the human, machine and natural domains. Horizontal relations denote pragmatic relations within domains, as described above and in Table 3.1. This is the case where agents are consuming (interpreting) signs produced by other agents within the same domain. In contrast, vertical pragmatic mappings apply between domains, where agents in one domain are interpreting signs from another domain—hence, humans and machines interpret natural signs as well as signs produced by each other.

The scientific knowledge cycle in Figure 3.2 further refines the relationships between the human and machine domains, represented by the inner and outer circles respectively. The natural domain is not shown, but is implied as being accessible to both humans and machines. Note that the cyclical representation explicitly recognizes the influence of prior knowledge, agent intention and action on the discovery of data, information and knowledge. The inner circle in Figure 3.2 shows a human-driven view of how knowledge is discovered and evolved via different inference types: empirical regularities typically denoting the properties of some phenomena are induced from observed data, theoretical elements are abduced from the regularities, predictions about unobserved world objects are made from the theoretical elements, thus building more complete (but hypothetical) possible models of the world, and these are tested against the real world, leading to verified model aspects, new observations and inductions, and so on.
The outer circle shows the mapping onto a machine-driven cyberinfrastructure: the mechanisms for inference (i.e. induction, abduction, deduction) are supported by computer operations, the mechanism for observation (action) is supported by instrument operations (humans or machines), and the inputs and outputs to both represent scientific artefacts stored in the system. The artefacts are categorized as data, information and knowledge. Data refers to the raw digital encoding of an observation from a human or instrument, and syntax to the grammatical and lexical conventions employed. Information refers to the partial models consisting of observed properties, typically stored in a repository such a database or even a web page, schematics refers to the conventions used for the structuring the repository, and signatics to the content of the repository. Theoretical knowledge refers to abstractions such as concepts, theories, and ontologies that can apply across geospace-time situations, and semantics refers to the abstraction conventions. Situational knowledge refers to more complete possible models of the world that contain referents to entities situated in geospace-time, some of which are inferred from theoretical knowledge. Pragmatics then refers to the origins, effects and uses inherent in all discovery operations, including a possible recounting of the motivations, goals, etc. involved.

The following is a simple example of theoretical knowledge discovery that uses the notation of machine or natural pragmatics as described in Table 3.1. The example assumes:
- concept formation operation \( Y \), input properties \( Z \), output concept \( X \), such that \( Y(Z, X) \);
- examples of \( X \) in the world called \( E \), exemplification function \( I \) for using \( X \) to get \( E \), such that \( I(X, E) \).

Then, informally, the following pragmatic aspects can be derived:

\[
\begin{align*}
\text{Eq. 3.1:} & \quad \text{Origin of } X = \{ \{ Y, Z \} \} \\
\text{Eq. 3.2:} & \quad \text{Effect of } X = \{ \{ I, E \} \} \\
\text{Eq. 3.3:} & \quad \text{Use of } X = \{ \{ Y, I \} \}
\end{align*}
\]

The application of this to situated geological regions is relatively straightforward, as shown in Figure 3.3: \( X \) denotes the geological concepts as described in a legend; \( E \) denotes polygons on a map that exemplify the geological regions concepts in \( X \); \( Z \) denotes the observed properties used to derive \( X \); \( Y \) denotes the learning procedure used to derive \( X \) from \( Z \); and \( I \) denotes the classification procedure used to label the polygons as members of \( X \).

![Diagram](image_url)

Figure 3.3: pragmatics for a geological map.
3.2 The nature of geo-pragmatics

The need for the knowledge cycle, and geo-pragmatics, is quite profound in the geosciences because models are often inferred from detectable properties and causal links. Such inference is necessary when the object as a whole typically cannot be fully observed due to its size and geospatial or temporal inaccessibility. For example, objects at a regional scale, sometimes called the geographical scale, cannot be viewed from one vantage point but must be interpreted from their traversal (Freundschuh & Egenhofer, 1997); and significant geologic objects may exist in the past or have parts hidden underground. Thus, some geoscientific objects may be less tangible because they are:

- historical artefacts, such as past phenomena whose residual effects still exert causal influence on the present, e.g. extinct volcanoes whose lava is in the rock record;
- interpreted objects, such as the objects discretized from continuous data;
- sampled artefacts, that are partially or indirectly detectable, or detectable at an inappropriate level of detail or scale.

Then, because the information used in such inferences is usually incomplete (due to sampling constraints) and the samples often heterogeneous (due to physical or social complexity), the concepts corresponding to such objects are also often incomplete and vague. This leads to competing models and ontologies on the one hand (e.g. Bie & Beckett, 1973), and on the other hand draws particular attention to the process of their discovery and evolution, and hence to the particularly geo-pragmatic aspects that frame their meaning.

This further draws attention to the need for a theory of geographical concept development that differentiates between direct and indirect apprehension of inferred objects. Direct apprehension implies an object is first identified and then its properties are apprehended, i.e. object identification precedes property detection. Concepts can thus be induced from clearly identified examples, including so-called “table-top” objects (Smith & Mark, 1998) or readily perceived geographic features such as buildings, ponds, etc (Mennis, 2003). However, the converse is true for indirect apprehension. In indirect apprehension the object as a whole is inaccessible, causing its properties and possibly parts to be detected first and the object and associated concept to be inferred subsequently. Examples of the latter include not only geographic regions such as geologic units, ecologic units, etc., but also macro-scale and micro-scale objects such as galaxies or electrons.

Therefore, geographic concepts can be induced from directly apprehended objects and their properties in unison, or both geographic concepts and objects can be individually abduced from detected properties and parts. This underscores the significance of the pragmatic aspects to both geographic concepts and objects. Being able to pragmatically explain a concept and object then becomes quite important in that the explanation enables tracking and reassessment of scientific artefacts. If our knowledge systems simply state the results of science (semantics) and forgo representing the process (pragmatics) then scientific replication and evaluation would be impossible, or at least severely limited. These issues are highlighted in Figure 3.4, Figure 3.5, and Figure 3.6.
The information and ontological squares shown in Figure 3.4 and Figure 3.5 respectively are neutral with respect to scientific process (Barwise & Seligman, 1997; Neuhaus, et al., 2004). The information square syntactically describes the logic of information flow from one combination of types and tokens to another, without denoting what these elements might be semantically or what the flow may mean pragmatically. The ontologic square goes one step further in that it semantically identifies its types and tokens as universals and particulars (i.e. abstractions and world entities) respectively, but without due regard to the process of abstraction—i.e. its arrows do not reflect a flow, or mapping, but reflect a relation. The question of how the relation comes to be is simply out of scope. This is a perfectly acceptable
approach ontologically, but one that falls short in supporting science where recording such pragmatic distinctions is valuable, even necessary.

The pragmatic square depicted in Figure 3.6 is proposed here to address these shortcomings. It builds on both the knowledge cycle and the ontological square. As shown in Table 3.2, the pragmatic square borrows inference claims from the knowledge cycle and provides associated pragmatic mappings and functions to explain how objects can be directly and indirectly apprehended from properties. Following the knowledge cycle, observed properties can be clustered into regularities, denoting concepts for properties, these can be abduced into concepts for objects, which can then deductively classify objects and instantiate object properties, where the objects themselves are individuated from other objects. A detailed example of this cycle is provided below in section 3.5.

<table>
<thead>
<tr>
<th>Pragmatic Mapping</th>
<th>Pragmatic Function</th>
<th>Source Entity Type</th>
<th>Target Entity Type</th>
<th>Entity Collection</th>
<th>Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensional</td>
<td>$F, G$</td>
<td>clustering</td>
<td>individual</td>
<td>concept cluster</td>
<td>Induction</td>
</tr>
<tr>
<td></td>
<td>$f$</td>
<td>blending</td>
<td>concept</td>
<td>concept blend</td>
<td>Abduction</td>
</tr>
<tr>
<td>Extensional</td>
<td>$g$</td>
<td>individuation</td>
<td>individual</td>
<td>concept situation</td>
<td>Abduction</td>
</tr>
<tr>
<td></td>
<td>$F', G'$</td>
<td>instantiation</td>
<td>concept</td>
<td>individual attribution</td>
<td>Deduction</td>
</tr>
<tr>
<td></td>
<td>$F'^{-1}, G'^{-1}$</td>
<td>classification</td>
<td>individual</td>
<td>concept class</td>
<td>Deduction</td>
</tr>
</tbody>
</table>

Entities in the pragmatic square that affect concept development are denoted as intensional, and entities that affect individual development are denoted as extensional. Collections of inputs or outputs to these functions and mappings are named clusters, blends, situations, classes and attributions, and are described further below. These functions apply strictly to scientific and machine pragmatics, as natural pragmatics solely involves natural functions.

The pragmatic square also borrows from the ontologic square the quality-substance distinction, between an entity and the properties it bears, but renames these to property-object only to underline the difference between the scientific artefacts of interest here, i.e. properties and objects, and their corresponding entities in the world. It also differs from the ontological square in other important ways:

- The elements are all scientific artefacts: concepts, individuals and mappings. Concepts refer to abstractions of properties or the objects that bear them—they are named in this way to differentiate them from their respective ontologic analogues of qualities and substances. Individuals refer to entities from which concepts are abstracted and the mappings denote discovery functions as described above in the knowledge cycle. This notion of individual is quite loose, referring in the short-term and purely for convenience to something akin to a specialization, i.e. any entity over which abstractions can be made. Individuals are defined much more precisely below, and in chapter 6, as singular objects situated in geospace-time—thus they will refer to a specific type of entity over which abstractions are made.
The linkages are mappings not relations: this reinforces the notion that knowledge is progressively discovered from existing knowledge, via inference and methodology, and contrasts with the ontological square that ignores discovery. It is important to note that the pragmatic square also presupposes the ontologic square in the sense that the mappings in the pragmatic square are used to create and revise the entities and relations in the ontologic square. But, the entities involved in the pragmatic square are only those related to discovery and revision, i.e. \( f(y_c) \) involves only those property concepts \( y_c \) used to infer or revise the target object concept, regardless of the object concepts with which they hold ontologic relations. In essence, the pragmatic square shows pragmatic mappings not ontologic relations (i.e. mappings related to origins, effects or uses), between properties, objects, concepts and individuals. Consequently, objects in the pragmatic square are assumed to hold separate binding relations to their properties once the objects are pragmatically discovered; so, if the pragmatic square is used to indicate \( G(x_i) = x_c \) for inducing an object concept \( x_c \) from object individuals \( x_i \), then the relations that bind the object concept to its properties are implied, i.e. \( G(x_i) \rightarrow R(G(x_i), y_c) \) where \( R \) is a binding relation between an object and a collection of some properties \( y_c \). In general, for property individuals \( y_i \), property concepts \( y_c \), object individuals \( x_i \), objects concepts \( x_c \), binding relation \( R \) and \( \text{instanceOf} \) relation \( I \) between a concept and one of its instances:

\[
\begin{align*}
\text{Eq. 3.4:} & \quad f(y_c) \rightarrow R(f(y_c), p_c) \\
\text{Eq. 3.5:} & \quad g(y_i) \rightarrow R(g(y_i), p_i) \\
\text{Eq. 3.6:} & \quad G(x_i) \rightarrow R(G(x_i), p_c) \\
\text{Eq. 3.7:} & \quad G'(x_c) \rightarrow R(G'(x_c), p_i) \\
\text{Eq. 3.8:} & \quad G'^{-1}(x_c) \rightarrow I(G'^{-1}(x_c), x_i) \\
\text{Eq. 3.9:} & \quad F'^{-1}(y_i) \rightarrow I(F'^{-1}(y_i), y_i)
\end{align*}
\]

The pragmatic square also particularly responds to the challenge posed by direct and indirect apprehension of objects. Development of object concepts from direct apprehension is accounted for by the pragmatic mapping \( G(x_i) \). Development of object concepts from indirect apprehension is accounted for by the pragmatic mapping \( f(F(y_i)) \). Development of object individuals in indirect apprehension is accounted for in the pragmatic square by \( g(y_i) \) which first identifies the object individual from property individuals, and then by \( G'^{-1}(g(y_i)) \) which classifies the object individual, and finally by \( G'(g(y_i)) \) which fills in gaps in the object’s properties by instantiating them from the concept. These development functions are summarized in Table 3.3.

<table>
<thead>
<tr>
<th>Table 3.3: direct and indirect of inference of objects in the pragmatic square.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Concept development (objects)</strong></td>
</tr>
<tr>
<td>----------------------------------</td>
</tr>
<tr>
<td>( G(x_i) )</td>
</tr>
<tr>
<td><strong>Individual development (objects)</strong></td>
</tr>
</tbody>
</table>

53
The presence of origins and effects: this reinforces the notion that the entities are connected within the geoscientific discovery process. The pragmatic square especially reflects this as it allows origins to develop from (a) a subset of all the entities that exemplify a concept, (b) example and contrast entities used to individuate the concept, or individual, from others or (c) instances of a concept’s properties. It also shows that concepts have effects in the form of classified individuals that are attributed properties from the concept or individual.

3.3 Elements of Geo-pragmatics

Although all three of the origin, effect and use aspects are important, this section will largely focus on the first two, origin and effect. The key elements are concepts, individuals, states, perspectives, versions, theories, ontologies, functions, clusters, blends, situations, classes and contexts, and all these are in turn described below.

Concepts

To summarize from chapter 2, concepts are defined here to refer to the abstract scientific artefacts generated via scientific method. They are neither mental entities, nor representations of patterns in nature called universals or kinds, but rather, they are scientifically accredited understandings of such natural patterns. This contrasts with a syntactic view of concepts that effectively associates them with a unary predicate symbol, and with the prevalent semantic view, which associates them either with a natural or mental artefact. The pragmatic position advocated here is aligned here with the claim that scientific knowledge lies in the relation between mind and reality, and this relation encompasses the scientific process of how reality comes to be understood (after Millikan, 2000). Under this view, understanding of reality is then, firstly, at best partial, as it is limited to what is detectable and testable by our sensory and instrument capacities, and secondly, it is a useful model of reality when scientifically accredited (Oreskes, 1994). Scientific artefacts may thus be variably useful representations of a portion of reality, i.e. ‘TRUE’, but this incompleteness implies there is more to discover. New methods and instruments, for accessing reality, or new samples, can then lead to augmented understanding.

This raises the issue of conceptual stability. Scientific concepts that are well validated (i.e. successfully tested against nature) are often thought to be stable and independent of mental or scientific process. This argument needs to be qualified in two ways: (a) scientific concepts evolve and are particularly unstable during knowledge discovery (e.g. discovery of the theory of plate tectonics, Thagard & Nowak, 1990), and (b) theoretical concepts are more stable than situational concepts, because of different levels of abstraction—the former are universal and can potentially apply to all geospace-time regions, whereas it is proposed herein that the latter are local to one space-time region and history, thus sensitive to changes in the underlying region or our sampling of it. For example, ‘Baker Brook basalt’ is situated because it refers to a natural grouping of sequenced rock layers found in possibly many rock bodies in a specific geographic area, whereas ‘formation’ is theoretical as it refers to the generally applicable conditions for rock bodies to be classified as formations. Now, ‘Baker Brook basalt’ may require conceptual revision if (a) new data reveals new properties, or (b) its members undergo a sudden geologic event that alters its properties (i.e. an addition to its history), but neither of these is likely to affect the ‘formation’ concept which would require a major scientific shift, perhaps one that is paradigmatic, to be dramatically altered. Thus
generalizations about a situation are more prone to conceptual change than generalizations over any situation. This distinction also serves to separate the concepts inherent in the historical and social sciences from others, as there is greater emphasis on situational concepts in the former than in the latter.

**Theories and ontologies**

Syntactically, theories often refer to a collection of well-formed expressions in a logical language, where well-formedness refers to conformance to a grammar defined by an alphabet and axioms for lexical composition. So, from a syntactic viewpoint, ontologies are then theories about concepts, that is, well-formed sentences defining or containing concepts. From a semantic viewpoint, ontologies are considered to be collections of specializations of *a priori* entities such as objects, processes, properties, roles, etc. (Guarino, 1998). But, as discussed above, the notion of situational concepts suggests refinement of the contents of domain ontologies to at least distinguish between situated and non-situated concepts. Furthermore, from a pragmatic viewpoint, if scientific theories, hence ontologies, refer to moderately proven scientific hypotheses that are not laws, because of residual uncertainties attached to them, then it becomes important to also capture the discovery and revision process—i.e. their origins—to aid in the evaluation of shifts in uncertainty of a theory. A good example of a situated theory might be the logical representation of a geological map legend, which describes the situated concepts that categorize particular objects on the map. A formal method of representing such origins is outlined further below.

**Individuals, states, versions and perspectives**

It is also important to distinguish between states, versions and perspectives of an entity. In formal ontology, a *state* refers to the unchanging qualities in some timeframe of a persistent entity (called an endurant or continuant (Grenon & Smith, 2004)). It is often characterized as a snapshot of the entity in some time interval that is produced by some event or process. This conceptualization is semantic, as it does not take into account the potential for multiple avenues of discovery for some state. For example, it is possible to describe the state of some natural object from fieldwork, lab analysis, or from more than one type of image classification, etc. Each such description can be quite different because methodological differences reveal different properties, while describing the same state in nature. The revealed properties can be disjoint or overlapping, as sometimes different methods uncover different properties (e.g. fieldwork and remote sensing) while other times the same property is uncovered by various methods (e.g. methods for obtaining the age of a rock). Hence, there is a fundamental difference between a state of nature and its various characterizations in methodological windows. Each such window is called a *perspective* here. The key notions are: (1) perspectives reveal various properties of a state and act as windows on it, and (2) perspectives are distinguished by methodology—by their discovery and evolution functions. Because the ontology for methodologies can be fine-grained or coarse, so can the types of perspectives. In some cases it is useful to have a remote sensing perspective, in other cases it might be useful to have perspectives for different types of image analysis method. The choice of the level of granularity will reflect the level of description desired for the state, such as one from a cumulative remote sensing perspective rather than from a specific imaging method, or one from a cumulative fieldwork view rather than from a specific field worker.
It is also possible to describe the same state in multiple ways within a perspective, e.g. multiple field workers may not agree on a description for the same state of an object, or a field worker’s opinion about a state might change (examples of both cases are shown experimentally in chapter 4). Likewise, different types of remote sense images input to the same classification method will likely yield different results for the state description, as might the same type of images taken at different times for an area. Each such variation within a perspective is called a version here. Unlike perspectives, which are demarcated methodologically, versions are distinguished by epistemic commitments to agents, intentions, or repetitions, i.e. different versions will likely be produced by different agents, or by different intentions in one or more agents (e.g. goals, purposes, beliefs, dispositions, etc.), as well as by repetitions of the same method by an agent who varies operating conditions. Under this framework an ontologic state, can be viewed as an abstraction generalized from one or more perspective states and any of these can be viewed as originating from possibly many version states. The interactions between these entities are illustrated in Figure 3.7. It shows the development of understanding over time for a rock body. The rock body as an ontologic individual and its ontologic states are shown in the topmost ovals. Several instances of two perspectives, one for fieldwork and another for remote sensing, are shown in rounded rectangles. These perspective instances show the rock body in four ways: from one map evolved over two time intervals (Map₁@tn, Map₂@tn), from a second map derived from the first map at a subsequent time (Map₂@tn₊₂), and from a remote sensed image obtained at a later time (Image₁@tn₊₃). Multiple versions of the rock body are developed at each time interval for the first map, and each version represents the view of a separate geologist (e.g. Geol₁@tn).
Vertical and horizontal relations denote generalization, evolution or verification. Generalization refers to development of an abstraction from less abstract entities, evolution refers to a revision of an existing entity, and verification implies confirmation of an entity across more than one perspective. Of note are the two types of generalization relations, one within a level (between version, perspective and ontologic instances of the state) and another between levels (between the ontologic individuals and states). This impacts the interpretation of the relations in terms of the pragmatic square: do they represent distinct pragmatic functions, such as clustering or blending functions, or do they represent an instance of the pragmatic square as a whole? Although the latter is probably the case, because each entity involves the assignment of properties to objects, in this dissertation it is only the inter-level generalization that will be modeled with the full pragmatic square; the intra-level generalization will be considered as denoting the blending pragmatic function because it operates within levels. This choice is mainly to do with the overall focus here on inter-level interactions and will require re-consideration in the future.

In the end, it is the lowermost versions of this hierarchy of states that are not treated as abstractions in this framework. Interestingly, it is such versions that are typically recorded in databases and that act as surrogates for ontologic states. This practice is likely related to the fact that in the geosciences a state of nature is typically sampled very few times; hence states, perspectives and versions are often conflated. But, when multiple methods are applied diversely by multiple agents, such as in complex multi-disciplinary ‘big’ science projects or even in multi-person mapping projects such as shown in Figure 3.7 above, it is useful to distinguish between ontologic states in nature, methodologic perspectives, and epistemic versions. One could then have, for example, a summary description of a state from fieldwork, based on multiple fieldworkers and supplemented by various lab analyses, a summary description from remote sensing, based on data from different sensors taken at different times, and a cumulative description of the state derived from multiple opinions and that reflects the best scientific notion about that state in nature. In such cases, the cumulative descriptions really represent socio-scientific products that result from various scientific interactions amongst agents. What is common and unchanging in states, perspectives and versions is the identity of the entity abstracted from them, i.e. states of an object have in common the object’s identity, perspectives of a state have in common the state’s identity, and versions of some perspective have in common the perspective’s identity and object’s identity.

The case is similar, but not identical, for concepts. In formal ontology concepts do not have perspectives and versions, as their discovery process is not considered. Yet it is evident that at least some concepts, such as situated concepts, also have perspectives and versions for the same reasons as states, i.e. methodologic and epistemic differences in their discovery and evolution. Thus one can speak of the different perspectives for the discovery of a situated rock unit concept, such as fieldwork, lab work, remote sensing, etc., and of different versions of a perspective resulting from (re)application of the discovery function by different agents with varying intentions. As with states, what is common amongst related versions is perspective identity, and additionally what is common amongst related concept perspectives is concept identity. Also similar is the notion that epistemic versions, methodologic perspectives and ontologic concepts are often conflated when the concepts are discovered from a single perspective and intention, and also the notion that they are distinguished in more complex scenarios.
A final consideration is the definition of individuals. So far, ‘individuals’ are implicitly defined as any entity over which abstractions can be made. This is a loose view of individuals in which individuals are the sources for abstractions. However, it is also possible in this view to regard individuals as targets of abstraction, derived from multiple source states at different time periods. The justification for this has to do with the role of discovery: ontologically it is sufficient to say that individuals have persistent identity over time, but pragmatics begs the question “how is identified conferred”? If identity and some persistent properties are built up as a result of a discovery process operating over multiple states and times, then individuals are indeed abstractions. Consider that ‘Boyan’ is an individual, that ‘Boyan at time $t_1$’ is a state of that individual, and that pragmatically ‘Boyan’ might be abstracted from various observed states. A more realistic example might involve a geologic region whose presence is inferred from current and historical states. Likewise, historical indicators of state often help identify other rock entities, such as when the states obtained during repeated seismic activity in an area indicate the presence of a fault. The key point is that state description can precede identification and thus serve as input to the abstraction of individuals. Clearly this is a pragmatic notion, because, ontologically, once the relation is made between an identified individual and state it is true for the full lifespan of the individual, whereas pragmatics considers how and when that relation is made.

What now remains to be determined is: (1) the type of discovery function involved in the abstraction of individuals, and (2) whether individuals have perspectives and versions. The answers lead from the assumption that an individual is an abstraction. Then, the discovery function must be intensional, because only the intensional clustering functions derive abstractions from less abstract entities. Also, then individuals are analogous to concepts and states, in that all can possess perspectives and versions. But there are differences between concepts and individuals: individuals are abstracted from one or more states over time, and concepts are abstracted from one or more individuals over space and time.

The abstraction process then becomes a complex affair as it not only involves mappings between abstraction levels, properties and objects, as shown in the pragmatic square, but also mappings between the ontologic, perspective and version views. Putting all this together results in an augmented pragmatic square, as shown in Figure 3.8, that accommodates concepts, individuals, states, perspectives and versions.

![Figure 3.8: augmented pragmatic square with perspectives, versions and mappings.](image)
The key aspects of this representation are the multiple vertical abstraction levels and horizontal partitions. Note the pragmatic square holds between vertical levels in a partition, and the abductive mapping $f$ applies horizontally within levels because it operates at the same level of abstraction. This implies it also applies to the abstractions within perspectives and versions, as well as between them—a temporary assumption made here mainly for reasons of scope. Later, in chapter 6, the concept level is expanded with more abstraction types, further expanding this model for abstractions.

A more expansive example of how the pragmatic square can be applied is show in Figure 3.9. The example illustrates the discovery of a situated geological formation concept (F1*), one individual (F1**) and one state (F11), from two perspectives: fieldwork and remote sensing.

Figure 3.9: example of the augmented pragmatic square.

The discovery process begins bottom-up, starting from two versions of the same state in each perspective. The fieldwork versions might reflect the descriptions of two geologists mapping
the same area, whereas the remote sensing versions might reflect different input image types such as Landsat ETM, Lidar or airborne magnetics, for the same area. The vertical abstraction relations (dashed closed arrows) illustrate abstraction chains within versions, perspectives and ontologies, whereas the horizontal relations (solid open arrows) show how cumulative abstractions are generated across versions and perspectives. It is clear from the diagram that in the absence of multiple perspectives and versions, the vertical groups will conflate to a single ontology.

In summary, the pragmatic approach semantically distinguishes entities in the horizontal direction as ontologic, perspective or version, and in the vertical direction as states, individuals and concepts. For the remainder of this chapter, and also for the most part in the next two chapters, individuals and states will be largely conflated. The main reason for this conflation is a primary focus on a single state, the current one, for any geoscientific object being considered. The two will be differentiated again in chapter 6. The pragmatic approach also distinguishes entities pragmatically by implying semantic relations from pragmatic mappings and their associated functions.

Scientific Reasoning
The reasoning constituents of scientific pragmatics are the induction, deduction and abduction inference mechanisms. C.S. Peirce originally describes these in two ways: according to their logical form (syntactically) and relative to their role in scientific practice (pragmatically), as elaborated in chapter 2 and summarized here.

Induction syntactically refers to the recognition of a pattern (Eq. 3.10), or pragmatically to the testing of a prediction. Deduction refers, syntactically, to reasoning forward from cause to effect (Eq. 3.11), or pragmatically to generating predictions. Abduction refers syntactically to reasoning from effect to cause (Eq. 3.12), the inverse of abduction, and pragmatically to the selection of the best theory based on prior knowledge such as analogous existing theories (Eq. 3.13). An example of the application of these reasoning forms is provided in the section below. The corresponding computational operations are summarized in (Gahegan & Brodaric, 2002).

\[
\text{Eq. 3.10:} \quad (A,B), (A,B) \vdash A \rightarrow B \\
\text{Eq. 3.11:} \quad a, A \rightarrow B \vdash b \\
\text{Eq. 3.12:} \quad b, A \rightarrow B \vdash a \\
\text{Eq. 3.13:} \quad (A,B), C \rightarrow D, A \approx C, B \approx D \vdash A \rightarrow B
\]

Clusters, clustering, intension and clusterMembership
A cluster is informally defined here to refer to the origin of a concept, where the concept is inductively abstracted from a collection of individuals. Because of the limited novelty provided by induction, a cluster represents nothing more than a grouping of source individuals that share, or contrast, properties according to some similarity metric, and the concept represents nothing more than a centroid amongst these individuals and can take its description from a representative example or an averaged prototype. The cluster might thus contain individuals with properties in varying contrast to the common set, in order to demarcate the edges of the cluster, which could be indeterminate. The associated pragmatic function is called clustering, which refers to the process of grouping individuals according to
similarity relations between properties. In the example above, the cluster for some \( X_j = \{ \{ Y_j, Z_i \} \} \), where \( Y_j \) denotes the clustering function used to induce \( X_j \) from some \( Z_i \). Note that the pragmatic square implies two types of clustering: clustering of properties, and clustering of objects in which properties are inherent. The first case results in a grouping of properties, and the second case results in some grouping of objects based on a grouping of their properties.

There are two associated relations, \textit{intension} and \textit{clusterMembership}. An intension is a relation between the induced concept and its associated cluster; it is a total and functional relation, meaning that each concept has a unique origin that, however, might be multi-modal in its distribution of properties. This also implies clusters can exist without being assigned concepts, a common notion in many literatures. ClusterMembership is a surjective relation between a cluster and an individual, indicating each cluster has at least one individual as a member.

The term “cluster” is chosen in part because of its use in spatial statistics, where it denotes a concentration of entities usually in geographic space. However, in this case it denotes a concentration of a concept’s properties in thematic space. The presence of instances as well as contrasts in the cluster is in agreement with the notion that some mental concepts are defined both proximal and distant examples (e.g. Tversky, 1977; Gardenfors, 2000).

**Blends, blending, intension and blendMembership**

A \textit{blend} is informally defined here to be the origin for a concept, and consists of a collection of source concepts that represent properties, analogies, background concepts, or contrast concepts. The associated pragmatic function is called \textit{blending}. In the example above, a blending for \( X \) is not specified, though it might be implied as being: the blend for some \( X_j = \{ \{ C_f, X_i \} \} \), where \( X_j \neq X_i \) and \( C_f \) is a blending function. The \textit{intension} relation also links the abducted concept to its blend—in effect it serves to relate a concept with its origins, those being either a blend or cluster. Then \textit{blendMembership} is a surjective relation between a blend and a concept, indicating each blend has at least one concept as a member.

Blending and clustering are distinguished according to: (1) their novelty, and (2) their source of abstraction. The first distinction results from the associated inference mechanisms: clustering is inductive and hence only mildly innovative as it represents a prototypical subspace of the input properties; in contrast, blending is innovative as the abducted concept contains properties, or their relations, not found in the input concepts. The second distinction is ontologic rather than inferential: clustering finds patterns in individuals, whereas blending develops new concepts from existing concepts.

The development of geological rock formation (object) concepts is an example of blending: prototypical property concepts might first be abstracted via clustering, and these property concepts might then be combined into a new formation (object) concept using prior knowledge of formation theory and analogies to similar formations in other areas. Another example is image analysis where spectral concepts are distinguished from information concepts: spectral concepts are essentially perspectival, as they are expressed in terms of the fixed properties of the image, whereas information concepts are essentially ontological as they are typically expressed in terms of domain properties. Clustering thus refers to the development of spectral concepts consisting of spectral properties, whereas blending refers to the development of information concepts in which the spectral properties are combined with
other perspective knowledge, e.g. fieldwork, to reveal a new ontologic situated (object) concept such as a geological formation.

The term “blending” is taken from (Fauconnier & Turner, 2002) where it refers to conceptual innovation via conceptual fusion. “Blending” in general is used here as a blanket term for a variety of cognitive integration mechanisms that lead to concept innovation, including conceptual combination (Wisniewski, 1998), metaphors (Lakoff & Johnson, 1980), image schemata (Lakoff, 1987), situational categories (Barsalou, 1999), analogies (Holyoak, et. al, 2001), and others, many of which have been shown to underpin scientific discovery (Feist & Gorman, 1998; Shrager & Langley, 1990; Zimmerman, 2000).

**Situations, individuation, situate and situationMembership**

A *situation* is defined here to refer to some context for an individual, and consists of a collection of other individuals that for some reason serve to contextualize the individual. Because a situation consists of some individuals only, it is in fact a partial model of the total collection of individuals. When the situation specifically denotes an individual’s origin, it then refers to the collection of individuals used to identify or differentiate another individual through an *individuation* function. Although a situation is not specified in the example above, it might take this form: situation for some $E_j = \{ \text{Id}_f, Z_i \}$, where Id$_f$ is an individuation function and Z$_i$ denote properties, parts and contrasts used to identify $E_j$. Situations will primarily be used in the latter sense in this dissertation, to refer to individual origins, keeping in mind that their full sense refers to a partial model of all individuals.

A prime example of individuation is the inference of a geological region from fragments of its boundaries. The primary role of individuation is identification, that is, differentiating new individuals. It is argued here that individuation is abductive because: (1) it is creative, i.e. a new individual is hypothesized from a series of ‘clues’, and (2) by process of elimination, i.e. induction results in concepts, not individuals, and deduction is ruled out because individuation precedes both classification and instantiation (Millikan, 2000), though in some cases it may be almost simultaneous with them—i.e. an object must first be recognized, however briefly, before it is classified or instantiated. Individuation implies a semantic relation called *sitruste* between an individual and its originating situation; it is a total and functional relation, meaning that each individual has a unique situation as origin, but a situation can individuate multiple individuals. Then *situationMembership* is a surjective relation between a situation and an individual, indicating each situation has at least one individual as a member.

Like other types of origins, i.e. clusters and blends, situations are significant not only theoretically, in support of enhanced meaning, but also practically in providing the raw materials to enable scientific knowledge evolution via evaluation, replication and testing.

Other accounts of situations in the AI or logic domains take semantic or syntactic approaches. For example, in situation semantics they are treated syntactically as the evaluation of a proposition, and semantically as a space-time slice of a possible world (Barwise & Perry, 1983); the latter dovetails with the full sense for situations defined above. The notion of situations as supportive of discovery (or abstraction) is rare, an exception being in situated cognition where situations are viewed as sensory inputs that lead to the establishment of mental concepts, and are vital to their subsequent mental usage (Barsalou,
These latter types of situations, that support discovery, can be viewed as a special case of the semantic notion of situations.

**Attributions, instantiation, instantiate and attributionMembership**

An *attribution* is defined as the collection of individuals whose properties have been revised or augmented via instantiation from a particular concept. This notion of instantiation is highly pragmatic as it primarily involves a revision function, i.e. the assignment of properties from a concept to an individual. Instantiation is typically used to fill empirical gaps in the properties of an object and thus represents an Effect of a concept. This differs from the semantic notion of instantiation, which generally refers to the relation between an individual and a concept with disregard to how that relation is produced (e.g. Grenon & Smith, 2004). Here, the *instantiate* relation is defined as the link between an attribution and the concept responsible for modifying the properties of the individuals in the attribution; it is a functional, surjective and injective relation, meaning that each concept may be related to one attribution and that each attribution is related to a unique concept. Then *attributionMembership* is a surjective relation between an attribution and an individual, indicating each attribution has at least one individual as a member.

Note that the transfer of properties from a concept to individual may be complex, because the properties of the concept and individual may not be identical but only similar, e.g. as in the prototypical concepts described in cognitive science (Gardenfors, 2000). Instantiation here also could then resemble the cognitive *simulation* function (Barasalou, 1999), in which the transfer process involves the selection and/or modification of pertinent concept properties based on the fit of prior situations to the current context of the individual. In this dissertation, such cognitive situations correspond to a concept’s origins, including its perspectives; instantiation then could involve fitting the current circumstance (or model) to a perspective in a concept’s origins, and adjusting the property transfer accordingly to deal with the degree of fit of origins to the current context.

**Classes, instances, extension and instanceOf**

A *class* is defined here, as per convention, as the collection of individuals exemplifying a concept, and an *instance* is defined as a member of a class placed therein via a *classification* function. So, in the example above, the class for some \( X_j = \{ \{ Cl_i, E_i \} \} \), where \( Cl_i \) is a classification function and \( E_i \) denote its instances. Classes represent an effect of a concept, because the concept must be known beforehand, prior to its use as an input to a classification function. The classification function is itself deductive because membership in a class is entailed when individuals pass classification criteria for the concept. Also implied is an *extension* relation that connects a concept with the class exemplifying the concept (Bittner *et al.*, 2004); it is a functional, subjective and injective relation, meaning that each concept may be related to one class and that each class is related to a unique concept. Then *instanceOf* is a surjective relation between a class and an individual, indicating each class has at least one individual as a member.

**Contexts and interpretations**

A *context* is informally defined here as the collection of origins, effects and uses for a concept or individual. From the example above, a context for some concept \( X_j \) can be determined: \( X_j = \{ \{ Y_j, Z_i \}, \{ Cl_i, E_i \} \} \). Contexts can be recursive such that they can...
additionally include the contexts for the various elements in some context. This leads to a spiraling collection that is potentially infinite in an open system, e.g. Context for $X_j = \{ \{Y_j, Z_i\}, \{C_l, E_i\}\}, \{\text{Context for } Z_i\}, \{\text{Context for } E_i\}\}$. To restrict this expansion, an interpretation is informally defined here as a single recursive path through a context (after Voissard, 1999), meaning that at each level only one element is chosen for the next stage of recursion, e.g. Context for $X_j = \{ \{Y_j, Z_i\}, \{C_l, E_i\}\}, \{\text{Context for } Z_i\}, \{\text{Context for } E_i\}\}$. This gives a limited, but useful, view of the causal chain of origins and effects for some entity. It can be used to isolate different perspectives or versions for an entity. Both contexts and interpretations are related to a notion of explanation found in computational reasoning, in which an explanation can be seen as a proof for an inferred conclusion where the proof consists of the chain of inferences leading to any logical conclusion, not necessarily restricted to a concept (e.g. McGuinness & Pinheiro da Silva, 2004). Chapter 5 has examples of contexts and interpretations.

3.4 Formalism for geo-pragmatics

Key components of geoscientific pragmatics are formally defined below using a second order logic representation (because functions are quantified). The formalization is not complete as associated axioms are not included but left to future work, e.g. axioms that formalize the total and functional character of the intension relation. Standard logic symbols are used: quantifier symbols $\forall, \exists$, definition symbol $\equiv$, logical and symbol $\land$, logical or symbol $\lor$, logical implication symbol $\rightarrow$, and function result symbol $=$. Variables are lower-case letters, and predicates begin in upper-case.

Primitives:
- $\text{Concept} \ (x)$ denotes $x$ is a concept;
- $\text{Individual} \ (x)$ denotes $x$ is an individual;
- $\text{Clustering} \ (f)$ denotes $f$ is a clustering function for inducing concepts from individuals;
- $\text{Blending} \ (f)$ denotes $f$ is a blending function for fusing concepts from other concepts.
- $\text{Classification} \ (f)$ denotes $f$ is a classification function used to classify individuals by placing the individual into class of an associated concept;
- $\text{Individuation} \ (f)$ denotes $f$ is an individuation function used to identify an individual from other individuals;
- $\text{Instantiation} \ (f)$ denotes $f$ is an instantiation function used to add or revise an individual’s properties from a concept;
- $\text{Member} \ (x,y)$ denotes $y$ is a member of collection $x$;
- $\text{Extension} \ (x,y)$ relates concept $x$ to a collection $y$ of individuals that denote its class;
- $\text{Intension} \ (x,y)$ relates concept $x$ to a collection $y$ of concepts or individuals, its origin;
- $\text{Situate} \ (x,y)$ relates individual $x$ to a collection $y$ of individuals, its origin;
- $\text{Instantiate} \ (x,y)$ relates concept $x$ to collection $y$ of individuals instantiated from $x$. 
Definitions:

Eq. 3.14: \( Cluster \equivdef (\exists y,z,f) (\text{Member } (x,y) \land \text{Individual } (y) \land f(y)=x \land \text{Clustering } (f) \land \text{Intension } (z,x) \land \text{Concept } (z)) \)

Eq. 3.15: \( Blend \equivdef (\exists y,z,f) (\text{Member } (x,y) \land \text{Concept } (y) \land f(y)=x \land \text{Blending } (f) \land \text{Intension } (z,x) \land \text{Concept } (z)) \)

Eq. 3.16: \( Class \equivdef (\exists y,z,f) (\text{Member } (x,y) \land \text{Individual } (y) \land f(y)=x \land \text{Classification } (f) \land \text{Extension } (z,x) \land \text{Concept } (z)) \)

Eq. 3.17: \( Instance \equivdef (\exists y) (\text{Class } (y) \land \text{Member } (y,x)) \)

Eq. 3.18: \( Situation \equivdef (\exists y,z,f) (\text{Member } (x,y) \land \text{Individual } (y) \land f(y)=x \land \text{Individuation } (f) \land \text{Situate } (z,x) \land \text{Individual } (z)) \)

Eq. 3.19: \( Attribution \equivdef (\exists y,z,f) (\text{Member } (x,y) \land \text{Individual } (y) \land f(y)=x \land \text{Instantiation } (f) \land \text{Instantiate } (z,x) \land \text{Concept } (z)) \)

Eq. 3.20: \( InstanceOf \equivdef (\text{Class } (y) \land \text{Individual } (x) \land \text{Member } (y,x)) \)

Eq. 3.21: \( ClusterMembership \equivdef (\text{Cluster } (y) \land \text{Individual } (x) \land \text{Member } (y,x)) \)

Eq. 3.22: \( BlendMembership \equivdef (\text{Blend } (y) \land \text{Concept } (x) \land \text{Member } (y,x)) \)

Eq. 3.23: \( SituationMembership \equivdef (\text{Situation } (y) \land \text{Individual } (x) \land \text{Member } (y,x)) \)

Eq. 3.24: \( (\forall x) (\text{Cluster } (x) \rightarrow \text{ConceptOrigin } (x) \rightarrow \text{Origin } (x) \rightarrow \text{Context } (x)) \)

Eq. 3.25: \( (\forall x) (\text{Blend } (x) \rightarrow \text{ConceptOrigin } (x) \rightarrow \text{Origin } (x) \rightarrow \text{Context } (x)) \)

Eq. 3.26: \( (\forall x) (\text{Situation } (x) \rightarrow \text{IndividualOrigin } (x) \rightarrow \text{Origin } (x) \rightarrow \text{Context } (x)) \)

Eq. 3.27: \( (\forall x) (\text{Class } (x) \rightarrow \text{ConceptEffect } (x) \rightarrow \text{Effect } (x) \rightarrow \text{Context } (x)) \)

Eq. 3.28: \( (\forall x) (\text{Attribution } (x) \rightarrow \text{ConceptEffect } (x) \rightarrow \text{Effect } (x) \rightarrow \text{Context } (x)) \)

Eq. 3.29: \( (\forall x) (\text{Interpretation } (x) \rightarrow \text{Context } (x)) \)

3.5 Example of geo-pragmatics

Figure 3.10 illustrates a cartoon-like example of geoscientific pragmatics in which geologic formation concepts and individuals are developed from observations of their properties (spatial orientation, temporal location, rock composition), and boundaries. In Figure 3.10 the lower left interior oval represents the observational data from which the formation’s property concepts are induced, and the geospatial boundaries of its instances are identified. The entities outside the interior oval, but within the exterior lower left oval, denote the relevant contrast observations. For example, the contents might be observations of rock compositions that are different from those inside the oval. When such contrasts are geographically adjacent they can be used to infer a boundary for the formation. The whole exterior lower left oval...
denotes the cluster from which the formation’s abstracted properties are generalized—these abstractions usually denote the prototypical properties for a formation. The upper left area of Figure 3.10 contains an oval for the prototypical properties, and a cloud for existing geologic formation concepts—taken together these enable abduction of a new formation concept, an object concept, one typically shown on the legend of a map. Finally, knowing some boundaries and having the formation concept allows the geospatial position and thematic description, respectively, of formation individuals to be inferred and a model denoting their collection to be built for the area. A conceptual schema for representing these pragmatic elements is presented in chapter 5 and their formal representation is sketched in a logic fragment below.

**Figure 3.10: example of geoscientific reasoning in the pragmatic square.**

### 3.5.1 Logic-based geoscientific reasoning in the pragmatic square

Figure 3.11 sketches the steps involved in the inference of a hypothetical formation from field-based observations, following the pattern outlined in the example above. The steps in Figure 3.11, though quite valid, depict a rather simple geological case:

A. Properties and objects are observed: several rock descriptions called lithologies (L11, L12, L21, L22) and a boundary between two lithologies (B11), respectively. Two lithology concepts (L1*, L2*) are induced (I1, I2) from these data.

B. A new boundary individual (B12) is abduced (A1) from the existing boundary (B11) and contrasting lithology observations (L12, L22). A boundary concept (B1*) is induced from the two boundary individuals.
C. A formation concept (F1*) and a formation individual (F11) are abduced (A2, A3) from existing theory about the nature of formations and from the prior observed and inferred entities. The formation individual is placed in the formation concept’s class (D1).

Figure 3.11: logic-based example of geoscientific reasoning in the pragmatic square.

The encoding below captures this example using logic. The same logical symbols are used as above; in addition constants are in capitals.

Predicates:
SpaceTimeRelation(OVERLIES): geospatially over and temporally younger relation
Near (x, y): spatial proximity relation
Bounds (x, y): y is a boundary for x
Constitutes (x, y): y is a material substrate for x
InstanceOf (x, y): y is a member of the x’s class
LithologyPI (x): x is a rock type property of an individual
LithologyPC (x): x is a rock type property of a concept
LithBoundaryOI (x): x is an individual boundary object
LithBoundaryOC (x): x is a situated boundary concept
FormationOI (x): x is an individual formation object
FormationOC (x): x is a situated formation concept
Functions:
\[ \text{LithStratPosition}(a,b) = z \]: lithologies a, b are related by a space-time relation z

Axioms:
X1 indicates a necessary and sufficient condition for a relatively homogeneous lithostratigraphic formation: such formations are constituted by a primary lithology and have a boundary between that lithology and others, thus preserving stratigraphic sequence.

\[ \text{X1: Formation}(x) \iff \text{Lithology}(a) \land \text{Constitutes}(x,a) \land \text{LithBoundary}(y) \land \text{Bounds}(x,y) \]

X2 denotes necessary and sufficient conditions for the boundary between two lithologies, such that the space-time relation between them is preserved in non-exceptional cases.

\[ \text{X2: LithBoundary}(y) \iff \text{Lithology}(a) \land \text{Lithology}(b) \land \text{SpaceTimeRelation}(z) \land \text{LithStratPosition}(a,b) = z \]

Observations

O1: LithologyPI (L11), a rock type individual is observed.
O2: LithologyPI (L12), a rock type individual is observed.
O3: LithologyPI (L21), a rock type individual is observed.
O4: LithologyPI (L22), a rock type individual is observed.
O5: LithBoundary (B11) \rightarrow LithologyPI (L11) \land LithologyPI (L21) \land \text{LithStratPosition}(L11,L21) = \text{OVERLIES},
A boundary B11 between lithologies L11 and L21 is directly observed and the spatial relation is noted as being OVERLIES.
O6: LithologyPI (L12) \land LithologyPI (L22) \land \text{Near}(L12,L22)
A nearness relation between two other lithologies, L12 and L22, is observed in geospace.

Then:

Induction: situated lithology concepts are induced from similar property individuals (i.e. pragmatic mapping \( F \)).
\[
\text{I1: L11, L12} \carr \text{LithologyPC(L1*)}
\]
\[
\text{I2: L21, L22} \carr \text{LithologyPC(L2*)}
\]

Abduction: boundary is interpolated between two lithologies (i.e. pragmatic mapping \( g \))
\[
\text{A1: LithBoundaryOI(B12) carr B11,O6}
\]

Induction: boundary concept is induced from observed and abduced boudaries and certain lithologic relations are implied (i.e. pragmctic mapping \( G \))
\[
\text{I3: B11, B12} \carr (\text{LithBoundaryOC(B1*}_{zw}) \rightarrow \text{LithologyPC(L1*)}) \land
\]
Abduction: a formation concept and object are abduced (i.e. pragmatic mappings $f, g$)
A2: FormationOC(F1*$_{fw}$) \[ L1*, B1*, X1 \] /* $f$ */
A3: FormationOI(F11$_{fw}$) \[ L11, B11, X1 \] /* $g$ */

Deduction: F11 is classified as member of F1* (i.e. pragmatic mapping $G'$)
D1: F1*$_{fw}$, F11$_{fw}$ \[ \text{InstanceOf(F1*$_{fw}$, F11$_{fw}$)} \]

Then:
For the fieldwork perspective, some of the pragmatic elements that can be derived include:
- Origin $\text{F1*}_w = \{ \text{A2, \{L1*, B1*, X1\}} \}$
- Effect $\text{F1*}_w = \{ \text{D1, \{F1*$_{fw}$, F11$_{fw}$\}} \}$
- Blend $\text{F1*}_w = \text{Origin for F1*}_w$
- Class $\text{F1*}_w = \text{Effect for F1*}_w$

Cluster $\text{L1*} = \{ \text{I1, \{L11, L12\}} \}$
Cluster $\text{B1*}_w = \{ \text{I3, \{B11, B12\}} \}$

Origin $\text{F11}_w = \{ \{\text{A3, \{L11, B11, X1\}} \}$
Situation $\text{F11}_w = \text{Origin F11}_w$

Now, suppose the same region is subsequently verified from remote-sensed images (i.e. machine origins). This introduces a new perspective in addition to fieldwork, and because only a single version is considered, perspective and version are conflated:
- $\text{Ii11}_rs$ denotes the collection of input images, such as hyperspectral images,
- $\text{Ai11}_rs$ and $\text{Pi11}_rs$ denote the agent and purpose (i.e. epistemic factors), such as the scientist analyzing the images and the motivation for the analysis, respectively,
- $\text{Clu11}_rs$ denotes a clustering function, such as an unsupervised or supervised inductive machine learning technique, that generates concepts and associated classification functions,
- $\text{F1*}_rs$ denotes a concept created by $\text{Clu11}_rs$ from $\text{Ii11}_rs$,
- $\text{Cla11}_rs$ denotes the classification function for $\text{F1*}_rs$, derived by $\text{Clu11}_rs$ from $\text{Ii11}_rs$,
- $\text{Io11}_rs$ denotes the collection of pixels placed in the class of $\text{F1*}_rs$ by $\text{Cla11}_rs$,
- $\text{Ind11}_rs$ denotes the individuation function in which a subset of $\text{Io11}_rs$ is identified as $\text{F11}_rs$: 

LithologyPC(L2*) $\land$ LithStratPosition(L1*, L2*) = OVERLIES
Cluster $F^*_rs = \{ \{Clu11_{rs}, \{Ii11_{rs}, Pi11_{rs}, Ai11_{rs}\}\}\} $

Class $F^*_rs = \{ \{Cla11_{rs}, \{Io11_{rs}, Pi11_{rs}, Ai11_{rs}\}\}\} $

Situation $F11_{rs} = \{ \{Ind11_{rs}, \{Io11_{rs}, Pi11_{rs}, Ai11_{rs}\}\}\} $

Then, for ontologic concept $F^*$, individual $F11$, blending function $Bld11$ which derives $F^*$ from $F^*_fw$ and $F^*_rs$, and individuation function $Ind11$ which derives $F11$ from $F11_{fw}$ and $F11_{rs}$, and classification function $Cla11$ which classifies $F11$ according to $F^*$.

Blend $F^* = \{ \{Bld11, \{F^*_fw, F^*_rs\}\}\} $

Situation $F11 = \{ \{Ind11, \{F11_{fw}, F11_{rs}\}\}\} $

Class $F^* = \{ \{Cla11, \{F11, F^*\}\}\} $ 

Furthermore, suppose that it is determined that $F11$ is caused by a local natural event $E11$, such as a sedimentation event, operating on an older rock body $R11$ identified as acting as one source for the sediments (i.e. natural origins), then a combined scientific, machine and natural origins for $F11$ is:

Origin $F11 = \text{Situation (}F11) =$ 
\[ \{ \{Ind11, \{F11_{fw}, F11_{rs}\}\}, \{E11, R11\}\} \]

Adding natural origins for a situated concept such as $F^*$ is analogous, e.g. suppose that $E1^*$ is a related regional event operating on several formation concepts $R1^*$ to produce $F^*$, then:

Origin $F^* = \{ \{Bld11, \{F^*_fw, F^*_rs\}\}, \{E1^*, R1^*\}\} $

And:

Context $F^* = \{ \{\{Bld11, \{F^*_fw, F^*_rs\}\}, \{E1^*, R1^*\}\}, \{\text{Origin }F^*_fw, \text{Origin }F^*_rs\}\}, \{\{\text{Cla11, \{F11, F^*\}\}}, \text{Class }F^*_fw, \text{Class }F^*_rs\}\} $

Interpretation $F^* = \{ \{\{Bld11, \{F^*_fw, F^*_rs\}\}, \{E11, R1^*\}, \text{Origin }F^*_fw\}, \{\{\text{Cla11, \{F11, F^*\}\}}, \text{Class }F^*_fw\}\} $ 

The interpretation here serves to showcase the fieldwork perspective by representing its abstraction chain.

3.6 Summary
This chapter makes a case for extending traditional meaning representation paradigms to include geoscience pragmatics, i.e. geo-pragmatics. It sketches a formal notion of geo-pragmatics and demonstrates its applicability via an example of geological formation discovery. The significance of the conceptualization is primarily theoretical—it theorizes the meta-functions and associated meta-entities involved in explaining geoscientific concepts,
and perhaps other geographical concepts. Its ultimate utility is in an enhanced representation of the concepts generated in the day-to-day practice of geoscience, as shown in chapter 5. In this way it responds to objective 2 of chapter 1, and geo-pragmatic challenges 1 and 2 from chapter 2, all of which call for the representation structure of concepts to be augmented with geo-pragmatic aspects.

Transposition of the example to a computational workflow is not demonstrated, but remains the subject of potential future work, as does consideration of how the pragmatic square applies to processes in addition to objects. Subsequent chapters will investigate the empirical basis for the conceptualization and example used in this chapter; they will also explore some practical and theoretical implications of the conceptualization.
CHAPTER 4: EXPERIMENTS TO ILLUSTRATE THE GEO-PRAGMATIC NATURE OF GEOSCIENCE INFORMATION

Previous chapters suggest that generalizations about geospace-time regions are a distinct type of concept, called a situated concept, and that such concepts are scientific artefacts whose meaning can be tied to geo-pragmatic aspects. This chapter looks for empirical evidence that this is the case. Specifically, it tests the hypothesis that natural and human origin artefacts are present in the information collected by some geoscientists collaboratively developing: (1) a map for a region, as well as (2) the associated situated concepts used to classify areas on the map. In the test, the field data of three geologists are compared and contrasted using geocomputational methods. Differences between and within individuals’ data for the same area would signify varied understanding of the concepts, and signal a need to better capture the geo-pragmatic situations in which the concepts were created and applied. Results are presented which corroborate this hypothesis, suggesting such concepts are influenced by natural, human, theoretical, and situational factors.

This study is generally different in three ways with respect to existing studies of geographic concepts (e.g. Smith & Mark, 1999): (1) it studies concept development rather than the character of established concepts; (2) it occurs in uncontrolled environments rather than in controlled laboratory situations; and (3) the scale of its subject matter differs, in that it studies concepts about geographic regions (i.e. geological units) that require a greater geo-pragmatic focus than concepts about geographic objects (e.g. mountain, lake) (for an exception to (2)-(3) see Suchan, 1998). This chapter responds to the third objective specified in chapter 1, and to the third geo-pragmatic challenge from chapter 2, both of which call for an empirical evaluation of geo-pragmatics.

4.1 Introduction

In traditional bedrock field mapping, geologists go to the field to develop a geological model for an area by interpreting field observations. The geologic model typically strives to include an explanatory geologic history for the area, described in terms of the three dimensional configuration of rocks, their temporal relationships and causal processes. The contribution of the interpretive element to geological reasoning is significant (e.g. Baker, 1999; Engelhardt and Zimmermann, 1982; Frodeman, 1995; Loudon, 2000; Martin 1998; Oreskes, et al., 1994; Schumm, 1991) but often difficult to identify and represent in non-human systems, as evident by the challenges encountered in modeling aspects of the geological reasoning process computationally (Burns & Remfy, 1976; Flewelling et al., 1992; McCammon, 1994; Simmons, 1983; 1988). The computational challenges posed by geological reasoning are related to geologists operating in open systems that provide few overarching constraints and that generally lead to multiple valid explanatory models as the field evidence and theory underdetermine the evolved space-time-process model (Martin 1998; Schumm, 1991). Constructing and selecting the optimal model is often regarded as part art, part science, and the artistry in geological mapping may well be related to the means by which explanatory models are constructed, and how a specific model is selected over others. Several factors may contribute to this construction and selection, including the prior knowledge brought to bear and an individual’s experience of the environment (Baker, 1999; Frodeman, 1995; Haugerud, 1998; Loudon, 1979), but little empirical work has been carried out to investigate these factors.
Field data are collected at sites and are typically represented as points on the map, whereas the interpreted regions that partition the map are typically represented as polygons. In addition, the polygons are classified according to a suite of situated concepts, the map units, normally described in the legend of the map. The field data thus represent, on the one hand, samples of the properties of the polygons, and on the other hand, instances of the prototypical property concepts used to characterize the map unit concept that classifies the polygons. E.g. the rock type observations for a group of polygons classified as map unit concept ‘X’ might include ‘granodiorite’, ‘monzogranite’, etc., which are thematically part of, or proximal to, the ‘granite’ concept. The prototypical rock type concept generalized from these observations might then be ‘granite’, which is used to characterize ‘X’. Thus ‘X’ may be said to be prototypically composed of granite, its core characteristic or prototype property.

To explore the nature of the situated map unit concepts and their relationship to the observations used to characterize them, this study compares the field observations to each other, and to the prototype. Degree of overlap between groups of field data related to various map unit concepts, as well as proximity of the groups to their prototypes, can then inform us about the nature of the map unit concepts and perhaps even about the process of discovery that gives rise to them.

One can imagine how useful such information might be: map features could be related to a specific point in the trajectory of a concept via a model of the development of the concept from the perspective of the map’s maker(s). Alternatively, tools that can describe the evolution of a concept through time could perhaps be used to show when one individual’s understanding of a concept has stabilized through repeated exposure, or alternatively where a group of scientists are recording the same observations but do not yet share the same understanding of what those observations imply. Such insights would be very useful for measuring the effectiveness of field science teams, for selecting ‘experts’ for particular tasks, for assessing consistency among individuals, and so forth. Similarly, tools that could visualise the supporting evidence during concept definition might lead to clearer insight and reduce disagreement between collaborators at the onset.

This chapter is structured as follows: section 4.1 provides a general introduction to the problem and methods employed; section 4.2 introduces the field area and the data collection methods, as well as some background material on concept formation; section 4.3 describes the methods used to correlate the field observations with each other and with the map unit concepts; section 4.4 presents the results and section 4.5 interprets the results; the chapter concludes with section 4.6.

### 4.1.1 Discovering situated geoscientific concepts

The interpretations in geological mapping differ somewhat from those in other forms of mapping in that geologists form situated concepts that classify past and present rock body individuals, while at the same time they construct a consistent geoscientific model to explain the evolution of those individuals in a region. The concepts are typically developed in-situ, often by teams of field scientists working together in a coordinated fashion, and they are not necessarily fixed at the outset, nor are they necessarily perfectly shared between all the geologists concerned. Rather, the concepts develop as a result of experience, exposure, training and other factors that are difficult, if not impossible to control. Because geoscientific environments are typically open systems characterized by compositional heterogeneity and
stochasticity, and because gaps in evidence and theory are abundant (Mann, 1993), and observations and interpretations can be somewhat subjective, it is possible for multiple models (i.e. maps) to validly explain an area (Schumm, 1991), even when concepts are shared.

This chapter sets out to study the development of geoscientific concepts in such environments in order to begin characterizing and evaluating geo-pragmatic knowledge. The logging records of individual field scientists are analyzed for consistencies and discrepancies. Because of the large number of variables and overall complexity of the data, geocomputational methods are employed for this analysis to gain insight into the nature of the concepts and their development through time. Previous work has shown that such concepts do not appear to be completely definable in terms of their properties alone, but in addition require significant knowledge that is not captured in the data (Brodaric et al., 2000). Needless to say, this apparent gap is detrimental to the effective and reliable use of the collected data and resultant interpretations. This gap is investigated to gain insight into the geoscientific concept development process and associated computing implications.

This chapter is also indirectly concerned with the question of whether semantic approaches to geospatial knowledge representation should be augmented by pragmatic considerations. Semantic approaches to geospatial information typically consider on the one hand, the types of universal concepts that are generally applicable (e.g. Grenon & Smith, 2004), and on the other hand, the perspectives in which specializations of such general concepts can be applied, such as the ecologic and geologic perspectives on a body of rock (e.g. Benslimane et al., 2000; Fonseca et al., 2000). However, versioning within perspectives according to epistemic factors is also important, and has practical value, because the same perspective could satisfy multiple models; i.e. geologists committed to a common fieldwork ontology might still retain individual versions of the ontology, potentially resulting in different maps (i.e. models) for the same area. To help explain such variation, previous chapters propose that concepts, and thus ontologic structure and perspective, are affected by the specific circumstances in which they are discovered and used. For information meaning, this implies that the meaning of a geospatial concept and individual might be contextual, in part due to ontologic commitments, and in part due to methodologic and epistemic considerations (i.e. how knowledge is acquired) involving the origins, effects and uses of the concepts. In the realm of geologic mapping this might translate into similar ontologic concepts in the same perspective being versioned differently by individuals, depending on the specific terrain encountered, the sequence of observation, their prior knowledge and expertise, as well as their methodological orientations, etc.

If pragmatic artifacts are indeed revealed by the data, then the need to weave pragmatic elements into knowledge representation frameworks, such as those discussed in chapter 3 and the next chapter, is substantiated. The main objectives of this chapter are thus aimed at understanding geological knowledge construction in field mapping in order to demonstrate the presence of pragmatic artefacts in the data, and to improve computational tools and methods that augment, rather than replace, geoscientists and existing practices.

4.1.2 Using SOM networks to study geoscience concept discovery

The interpretations developed in a geologic mapping exercise can be characterized as consisting of concepts and individuals, and the relations and functions between and within
them. The collection of concepts, their relations and functions is referred to here as a concept model, or ontology as defined in chapter 3, and the collection of individuals and their associated entities is here called an individual model. The concept model for a geologic map consists of a logical organization of theoretical and situated geologic concepts that detail a hypothesis about the geologic evolution of a geographic region; the individual model consists of concept instances such as representations of concrete physical individuals, relations and processes extruded in geographic space and geologic time. The distinction between these two types of models is often made for scientific information (e.g. Langely, 2000; Suppes, 1962).

As shown in Figure 4.1, an example of a concept model is the collection of conceptual components found in a geological legend and associated reports (minus the cartographic descriptions); these describe the concepts and relations used to classify the individuals represented within the geospatial extents of the map. The classified individuals, denoting rock bodies and their relations, then represent the individual model. Note again that individuals are conflated with states here. For geologic map units, the conceptual model primarily involves situated concepts, and the individual model primarily involves geological regions, their boundaries and associated relations. Importantly, a geologic map is deemed to work if the conceptual and individual models conform, without conflict. Then, a geological model can be informally defined as a ‘working’ combination of an ontology and individual model, and a geologic map can then be defined as a symbolized geological model represented in some medium (after Brodaric & Hastings, 2002).

![Figure 4.1: ontologies and individual models in geological mapping.](image)

The purpose of the chapter can now be stated more precisely as the study of the development of the concept model, and specifically, the concepts, blends, clusters and classes associated with it. Because clustering approaches are used, the study will focus on: (1) the nature of the
clusters containing observations of the properties of map unit instances, and (2) their relation to the prototypical property concepts that characterize the map unit (object) concepts. Clustering will therefore refer to the development of property concepts from field data, and clusters will refer to the collection of field data used to generate the property concepts. For shorthand, the terms ‘concept’ or ‘map unit concept’ will refer in this chapter to the combination of the map unit (object) concept and its generalized properties, such that in this shorthand clustering will refer to the development of both the generalized properties as well as its map unit concept. This admittedly omits the blending step by which the map unit (object) concept is generated from the generalized property concepts, but since limited insight into blending can be gotten from the data, the process of map unit discovery is conflated with clustering here, for convenience only. Clustering, or conceptualization, will thus refer here to the grouping of field data from which the generalized properties of a map unit concept are abstracted, and hence it will also refer to development of the map unit (object) concept itself.

The chapter explores the degree of positive association within and between clusters of the map unit concepts, using clustering metrics, thereby providing a measure of the degree to which such concepts are defined by their clusters and, conversely, also providing a measure of the lack of clustering—i.e. a confusion metric—indicating the degree to which: (1) factors other than the recorded data might influence clustering and/or (2) the concepts reflect heterogeneity in the physical environment. To the degree that clusters represent collections of field data that are critical to concept formation, these metrics can be viewed as indicators of the impact of clustering on concept contents, and may inform conceptual coherency, consistency, etc.

To examine clustering, the field data and concepts are first associated using two neural network techniques: the inductive self-organizing map (SOM) and learning vector quantization (LVQ) (Kohonen, 1997). These were selected because of their ability to cluster in high dimensional spaces, including their accommodation of complexities such as non-linearities and inter-dependencies, and their accessible clustering/confusion metric (for alternative measures of map concept confusion and uncertainty see Apsinall & Pearson, 1994; Bierkens & Burrough, 1993a,b; Goodchild et al., 1992). Both neural network approaches take as input the observed field data, converted to numeric vectors, and induce a reduced number of representative vectors denoting the generalized properties for the unit, and in our shorthand, thus also for the unit itself. The SOM derives these representative vectors strictly from the input field data without consideration of the derived map unit concepts, whereas in supervised clustering via LVQ the map unit concepts are also used. Each representative vector stands for field data that are more similar to it than any other representative vector, based on Euclidean distance. Each representative vector can also be labeled according to the map unit concepts of the field data it represents. Visualizing the labeled representative vectors then shows the spatial distribution of clusters without the visual clutter engendered by duplicate or similar vectors. Positive clustering (forthwith just “clustering”) is evident when same-labeled vectors form relatively homogenous spatial groupings, and confusion is evident when clusters do not form groups, when groups are heterogeneous, or when representative vectors possess multiple labels in the SOM. Essentially, confusion results when cluster members overlap, i.e. when similar field data are recorded for multiple map unit concepts, and clustering occurs when field data membership is unique, i.e. when the field data in one cluster do not occur in another.
In addition, clustering behavior is inspected over discrete time intervals using statistical techniques that measure the diameter of a cluster, or the distance between two cluster centers. Measuring these at discrete time intervals then produces graphs that can be used to illustrate: (1) geologist’s clustering patterns within map unit concepts over time, and (2) the similarity between geologist’s clusters over time. This should aid in the understanding of the development of the map unit concepts, and further inform conceptual consistency, coherency, etc., within and between geologists—it should provide insight into the degree of versioning occurring.

4.2 Background

4.2.1 Geological Study Area and Field Mapping Methodology

Six contiguous geological maps (Tella, et al., 2000) and supporting field evidence were obtained from a geological mapping project carried out in northern Canada (Figure 4.2) by the Geological Survey of Canada and its partners. The single legend for these 1:150,000 scale maps summarizes several situated concepts (e.g. map unit X, rock type for X) and theoretical concepts (e.g. fault), which comprise the conceptual model for the area. These are described more fully in related publications (e.g. Hanmer & Relf, 2000). Apart from describing various geological concepts in the legend, the planar maps also show the classes and clusters for these concepts consisting of individuals with polygon, line, or point geometries.

Figure 4.2: the study area in the Macquoid geological belt of northeastern Canada.

To gather the field data, 15 geologists of diverse expertise operated from a common base camp for varying lengths of time, criss-crossing the area on foot to obtain mainly site-based information from which full regional coverage of the area was interpolated and abstracted; but for the purposes of this study the data from the three geologists primarily concerned with developing the overall conceptual and individual models were selected for use. In total, these three geologists visited approximately 1500 field sites without repetition or overlap, with relatively equal distribution of sites per geologist, and fairly equal distribution of geologist sites per map unit concept (Figure 4.3, left). Concepts with larger spatial extents were more heavily sampled (Figure 4.3, right), almost every map polygon was sampled by at least one of the three geologists; moreover, a significant proportion of the polygons was sampled by 2 or more of the three geologists (though via different field sites). Hence, the geologists’ field data were evenly distributed both spatially and thematically.
The field data were collected and managed in situ with the aid of the FieldLog digital geological field system, which deploys hand-held computers and satellite-based positioning systems (GPS) and enforces a common but customizable data structure and vocabulary for all data recording and map construction activities (Brodaric, 2004). Geologists were able to systematically and consistently record data on-site into various tables. The tables contained fields for geographic position, date, geologist’s name, lithology description, structural orientation measurements, sample descriptions, photograph summaries, digital sketches, interpretive comments, and others. Significant effort was made to keep the tables, fields and vocabulary consistent between individuals as the fieldwork progressed and understanding evolved. Hence, not only was the spatial and thematic coverage of the area evenly distributed, but many aspects of the conceptual model, such as the various geoscientific terminology employed, were also coordinated between the geologists and kept consistent. Once collected, the data were regularly transferred from the hand-held systems to a central geospatial database operating on a generator-powered laptop at the base-camp. Elements of the database were then analyzed by the geologists and subsequently plotted to an evolving digital map, so as to guide future fieldwork, and coordinate individual understandings into a uniform interpretation. It is this database and the digital and paper versions of the resultant maps that were subsequently obtained and analyzed to gain insight into map unit concept development.

The database of field observations used in this study is ideal in many respects: it is well-structured and fairly representative of the domain; the sample size (number of sites visited per concept) is quite large by the standards of laboratory controlled experiments; the conceptualization process was explicitly coordinated to achieve harmonized concepts between participants; and, moreover, the data reflects in situ observations obtained without recourse to memory. Reliance on memory is often the case in controlled studies but it is not necessarily the preferred experimental methodology for studying scientific knowledge discovery (Dunbar, 1997; Klahr & Simon, 1999). The database is therefore well suited to the study.
4.2.2 Geological Uncertainty

One would expect the development of situated geoscientific concepts to be a complex process, characterized by the tension between what is known about some geological concept, cluster or class which constrains potential models, and what is unknown, which increases the number of potentially valid models. Moreover, we might also expect this tension to shift in emphasis from unknown to known, as understanding in an area evolves and as models are consequently refined or eliminated. Points in this range of tension may be seen as measures of uncertainty, and we might expect greater uncertainty to result in decreased concept understanding, whereas greater certainty would reflect increased understanding. Measures of confusion and clustering obtained from inductive analysis might then reasonably be used as indicators of uncertainty and certainty, respectively. Factors affecting uncertainty are therefore important to geological concept formation, as they help explain clustering and confusion in the field data, and these factors may in general be attributed to: (1) human aspects, including incomplete measurement or observation, due to data scarcity and human subjectivity; (2) physical aspects, related to the irregularity and complexity of the physical system; and (3) scientific aspects, involving imperfect or incomplete scientific knowledge often due to gaps in evidence or theory (after Mann, 1993).

Characterizing the behaviour of a concept’s cluster through time should thus provide insight into various uncertainties associated with the evolving concepts themselves. Related studies (Hunsaker, et al., 2001) have focused on: (1) uncertainty in the instances within a class, associated with the boundaries, spatial distribution and internal composition of mapped geospatial individuals (e.g. Bierkens & Burrough, 1993a; Gahegan & Ehlers, 2000); (2) uncertainty in the conceptual model, related to gaps in knowledge such as missing concepts and theory (Mann, 1993) or to the confusion between geoscientific clusters compared without consideration of their development in time (Aspinall & Pearson, 1994); and (3) uncertainty in the relationship between concept and class (Freska & Barkowsky, 1996). In this case study we concentrate on the second type of uncertainty, related to the conceptual model, by examining the development of a concept’s cluster through time.

4.2.3 Cognitive Models of Knowledge

The discussion above suggests uncertainty in geoscience arises from human, scientific and physical factors, and that these factors might explain levels of confusion, or clustering, in the inductive analysis of the field data. In this section it is further suggested that these factors might partially map, respectively, onto cognitive models of knowledge consisting of implicit, explicit, and situational knowledge, and that these comprise three significant aspects of geoscientific knowledge as used and developed during field mapping.

Implicit, Explicit and Situated Knowledge

Implicit knowledge is typically characterized as tacit, experiential, intuitive, unconscious, perceptual, visual, non-symbolic, and procedural (i.e. knowing-how), whereas explicit knowledge is portrayed as rational, logical, conscious, conceptual, verbal, symbolic, and declarative (i.e. knowing-what; Goschke, 1997). This distinction conforms to the split in many fields between intuition and logic and also perhaps between human and scientific uncertainty, e.g. in scientific method (Hanson 1958; Kuhn, 1962; Popper, 1959) or in artificial intelligence (Clancey, 1993; Vera & Simon, 1993). In field-based surveying these opposite poles might manifest themselves as implicit knowledge of pressure, temperature,
and composition as evidenced in observable properties of rocks and constituent minerals on the one hand, and as the germination of such experience into explicit interpretation and model construction on the other. Implicit knowledge may as such account for some of the ‘artistry’ in such transitions with explicit knowledge correspondingly accounting for ‘scientific’ aspects.

In spite of the obvious dichotomies, implicit and explicit knowledge are generally both thought to bear on a situation and thus influence its outcome. The model resulting from the integration of implicit and explicit knowledge in situations is often called a mental model (Bara, 1997), and is called situational knowledge. Because situational knowledge emphasizes how external circumstances interact with implicit and explicit knowledge to form understanding in the moment, we can say it acts to coordinate human, scientific and physical influences on knowledge acquisition, with some special emphasis placed, for field-based geoscience, on the contribution of immediate physical circumstance to this coordination. It effectively serves to bring together the ‘art’ and ‘science’ of mapping to a circumstance. Apart from this focus on immediate circumstance, the situational approach also emphasizes the development of knowledge over time, as it purports that human knowledge and our learning capacities are updated in every circumstance—in this view humans are inherently learners affected by the nature and sequence of the circumstances encountered, reinforcing the notion that data clusters and related concepts may evolve in time (Smith & Samuelson, 1997). In geological mapping this situational focus often expresses itself as knowledge-driven exploration, where the sensing of immediate geological conditions may invoke model re-evaluation, which in turn affects subsequent mapping including what is observed, where observation occurs, and how it is evaluated.

Two implications important to the analysis of geological clusters and concepts arise from consideration of types of geoscientific knowledge. Firstly, because implicit knowledge components are by definition not expressible they are impossible to fully capture in data and may thus augment confusion in the geocomputational techniques. Secondly, confusion may be related to explicit concept structures.

**Geoscientific Cluster Structures**

Two significant factors in explicit cluster structure, as discussed by cognitive scientists, are central tendency and grading (e.g. Barsalou, 1985; Medin, 1989). Central tendency refers to the level of clustering of characteristic attributes of the data for a concept—i.e. to what degree does the data tend to one or more core clusters? Grading refers to the level of typicality of a datum in a cluster—i.e. how typical is the example of the concept? Clusters have been empirically shown to generally possess central tendency and grading, but these two factors are not always connected, in that definitive examples for some concepts may exist outside the central tendency of its cluster: i.e. in data-driven concepts central tendency and grading align, whereas in knowledge-driven concepts they can divorce (see Figure 4.4).

In data-driven probabilistic clustering, grading and central tendency align, inasmuch as definitive examples of the concept are found in the central tendency of the data. Thus, concepts induced from data and their clusters are structured around an average, or central tendency, of the attributes of all observed examples such that examples closer to the central examples are more typical than examples more distant (Barsalou, 1985; Lynch et al., 2000)—e.g. one would expect data-driven probabilistic structure to be exhibited by a map
unit concept that has heterogeneous composition in which one or more compositional elements are clearly dominant.

Graded structure is also found in knowledge-driven theoretic conceptualization where concept examples also have varying typicality but where the definitive examples are not necessarily in the central tendency because assignment of the example to a concept is primarily reasoned from theory rather than induced from data. For example, concept structure may be organized around theoretically significant ideal attributes that occur outside the central tendency; e.g. data that are not often observed but are significant to a map unit concept, such as when the boundaries for a map unit are rarely observed relative to other compositional and structural data, or where there exists no definitive examples (i.e. type localities) for a map unit because it is an ideal abstraction or aggregation reasoned largely from prior theory due to data gaps, sampling issues, etc. Such concept structures therefore result from factors in addition to data, such as pre-conceived goals (Barsalou, 1985) or theories (Murphy, 1993). Knowledge-driven theoretic conceptualization is significant to geoscience in that concepts and individuals are often inferred from prior knowledge because they cannot be directly observed or measured.

Both data-driven and theoretic conceptualization may be associated with explicit knowledge insofar as the central tendency, ideals, and theories are expressible. Graded structure also characterizes concept development, and thus some situated knowledge, as data and theory may interdependently co-evolve in time (Wisniewski & Medin, 1994) causing examples to become marginal and fall outside a central tendency when conceptual and individual models are abandoned or revised.

Intuitively, stable concepts should trend to a core cluster representation, even though this core might (or might not) exist in a graded structure and possibly not fall within the central tendency. However, convergence to a core cluster representation is not guaranteed for all
concepts, as some concepts may be inherently unstable and exist in a constant state of flux, due to highly variable situations, inadequate sample sizes, or because of the inadequacy or instability of guiding theory and models (Smith & Samuelson, 1997). For example, generalizations over chaotic physical systems will result in unstable concepts that reflect the underlying shifts with such systems. The conceptual and individual models then generated are mainly descriptive, rather than explanatory, consisting mostly of the observations taken and the vocabulary employed, respectively. In essence, uncertainty is too high in such models to reach mature concept development or explanation.

The considerations above ramify the following for analysis of the field data. The clusters associated with situated geological concepts may be graded, exhibiting areas of varied clustering where confusion may be due to marginal examples that possibly arise from: (1) overlapping data-driven clusters, (2) the disconnect between examples in the central tendency and those theoretically significant, and (3) revised or discarded models and theories generated during concept development. Importantly, one would expect clusters to become reinforced in time, tilting the balance towards higher clustering and lower confusion, and to models and a map with greater explanatory power. But, high uncertainty caused by various human, physical and scientific factors may tilt the balance oppositely, toward higher confusion and lower clustering, and to models and a map that are more descriptive than explanatory—i.e. clustering in highly variable situations, unfamiliar terrain, or using small sample sets, may preclude explanation development. Owing to the prior knowledge that the models and maps developed for our study area are primarily explanatory in nature, rather than overtly descriptive, it is expected that any confusion in the results is probably be due to: (1) implicit influences, or (2) explicit influences involving overlapping, theoretically significant or discarded examples. The results in section 4.4 support such interpretation.

4.3 Representation and Visualization Methodology

Field-based concept development was investigated by building and visualizing neural network representations of the geological field data. Considerable effort was first undertaken to convert the largely textual geological observations to numeric vectors suitable for neural processing and representative of the original data. The inductive neural learning techniques of supervised clustering using Learning Vector Quantization (LVQ—Kohonen, 1995), and unsupervised clustering using the Self-organizing Map (SOM—Kohonen, 1995), were then utilized to construct a multi-dimensional feature space containing vectors prototypical of the input numeric field data vectors. Prototypical representation reduces clutter for visualization, emphasizes conceptual trends, and quantifies confusion, while being sensitive to frequency of occurrence of the original data. The prototypical vectors were subsequently labeled according to the concept of the nearest matching field data. Labeling occurred in two modes, before and after prototype construction in supervised clustering and unsupervised clustering, respectively, allowing for cross-referencing of results. Confusion is apparent when multiple labels were assigned to prototypes in unsupervised mode and was implied when labeled vectors did not cluster in both supervised and unsupervised modes. Accuracy values for a set of samples can be determined by applying the data to a trained SOM and recording the proportion of correctly identified samples.

The conceptual uncertainty and consistency of the concepts is also inferred by examining the development of geologic clusters through time. This is achieved by projecting the data into a multi-dimensional ‘feature’ space then measuring the approximate diameter of the resulting
clusters’ central tendency within this space per individual field-scientist over time, and measuring the approximate clustering distance between pairs of individuals’ through time. The technique is also used to determine geospatial clustering of the sites where the data are observed, and to measure the distance between the clusters of field data and the prototype cluster for each concept, taken from the map legend.

The clustering techniques and data preparation issues are described in the following sections.

### 4.3.1 Self-Organizing Map (SOM)

The unsupervised SOM consists of an initial two dimensional matrix of connected vectors, its nodes, each possessing a set of weights. In training, the numeric input vectors are repeatedly introduced to the matrix causing node weights to be adjusted within a locus around the node nearest the input vector via Euclidean distance. The amount of adjustment is calculated as the vector difference between the input vector and nearest node, and this amount is propagated to neighboring nodes (Figure 4.5). Both the adjustment amounts and propagation extents may be configured to decay over the course of training. Once trained, each node may be labeled by assuming the concept from the nearest original data vector. Nodes may be assigned multiple labels if they respond to multiple concepts, thereby providing a measure of confusion between clusters when the same datum is grouped into multiple clusters each corresponding to a unique concept. Because the aim of the SOM is to capture and reflect the topology of the input data, visual inspection of a SOM provides a cursory but effective indication of concept clustering and confusion. However, because the nodes are arranged at fixed intervals, the relative distances between nodes do not provide a precise view of relative node similarity. Applying the distance-preserving Sammon mapping (see below) to a SOM permits visualization of relative node distances and cluster similarity. The SOM is useful for visualization purposes inasmuch as it is a generalized view of the native distribution of the data in feature space.

![Figure 4.5: training a SOM: the nearest node (dark gray) is found and moved toward the input vector. A weighted adjustment is also propagated to surrounding nodes (light gray) within a certain radius.](image)

### 4.3.2 Learning Vector Quantization

Learning Vector Quantization (LVQ) extends the SOM for supervised learning. In LVQ, the labeled input data vectors are repeatedly presented to an unconnected set of nodes initialized with concept labels and statistically meaningful weights. During training the nearest node is found and adjusted toward the input vector if their labels match, and moved away from the input vector if the labels mismatch. Note that unlike the SOM, LVQ does not propagate adjustments to other nodes, though a time-decay learning rate may be defined to control the amount of adjustment applied per iteration. Also unlike the SOM, LVQ is biased towards...
cluster separation and is therefore a less accurate reflection of the native distribution of the data in feature space. LVQ is particularly useful in that it does provide a quantitative measure of similarity, or clustering and conversely confusion, between data and the neural representation in terms of the proportion of data correctly classified by the neural representation. LVQ nodes may also be visualized via the Sammon mapping discussed next.

4.3.3 Sammon Mapping

The Sammon mapping (Sammon, 1969) is a form of multi-dimensional scaling that transforms an \( n \) dimensional vector space into a two-dimensional plane suitable for visualization. It preserves relative vector distances but not topology, so the shape of original clusters may be deformed. This transformation enables cluster similarity and cluster presence to be viewed once the nodes are labeled, because similar vectors and hence clusters are nearer to each other. Of particular significance is the fact that overlapping clusters and multi-labeled nodes signify similar input data and thus indicate potential conceptual confusion.

4.3.4 Mean Median neighbour Distance (MMD)

To compute measures of clustering over time, the field evidence is first transformed into a linear numeric vector for each observed site. A clustering metric for a set of vectors \( (A) \) is obtained by calculating the median neighbor distance for each vector and then averaging them. So, if vector set \( A \) possesses \( k \) vectors, each with \( n \) attribute dimensions, then:

\[
A = \{v_1, v_2, ..., v_k\}, \quad v_i = (a_{i1}, a_{i2}, ..., a_{in}) \quad 1 \leq i \leq k.
\]

The Euclidean distance between \( v_i \) and \( v_j \) was calculated as:

\[
d_{ij} = \left( \sum_{m=1}^{n} (a_{mi} - a_{mj})^2 \right)^{1/2}.
\]

The clustering diameter for \( A \) is then formulated as the mean (average) of the median (M) of the distances \( d_{ij} \):

\[
D_A = \frac{1}{k} \sum_{i=1}^{k} M(d_{i1}, d_{i2}, ..., d_{ik}), \quad d_{ii} \text{ excluded.}
\]

This is a variation on a technique in which the median nearest neighbor distance is used to measure clustering in high dimensional spaces (Kohohen, 1997). However, the mean median neighbor distance (MMD) used here has the benefit of incorporating isolated clusters, unlike the nearest neighbor (Clark & Evans, 1954), without exaggerating them, like the mean neighbor. Taking the mean of the median distances, factors in potential multi-modal distributions among the median distances.

MMD can also be used to determine a similarity metric for two sets of vectors by measuring the distance between their central tendencies. Due to asymmetry, the distance between vector sets \( A \) and \( B \) is calculated by averaging the mean of the median distances from \( A \) to \( B \) and the mean of the median distances from \( B \) to \( A \): i.e. the left side of the equation takes the median distance from each vector in \( A \) (\( k \) vectors) to each vector in \( B \) (\( p \) vectors) and visa-versa for the right side; these two values are then averaged.
The impact of outliers can be modulated by varying the scaling of the vector dimensions, thus condensing or expanding the overall ‘feature’ space, and increasing or decreasing, respectively, the contribution of any single dimension to the overall clustering.

In this study, the key ‘dominant rock type’ dimension is emphasized in order to improve differentiation of clusters. Emphasis is achieved, firstly, by scaling the dominant rock type dimension so its values exceed those of other dimensions by several orders of magnitude, and secondly, by numerically encoding the taxonomic structure of the nominal rock type values such that successive levels use numbers smaller by a factor of 10, resulting in taxonomically closer values being nearer in feature space (Figure 4.6, right). Since taxonomically similar vectors are grouped nearer each other than taxonomically distant vectors, and because the taxonomically arranged dimension is critical to the differentiation of the clusters, the methods described above serve as a useful measure of the degree of clustering within a concept and also as an informative measure of the discrepancy between two sets of clusters.

Most of the remaining \( \approx 100 \) dimensions, such as specific fabrics, textures, colors, minerals, etc. are also scaled to ensure that, for example, all ‘granodiorite’ variations of a rock type are nearer each other and hence more similar than ‘monzogranite’ variations or other rock types (see Figure 4.6 left). Since these remaining dimensions are scaled to fall between the values of 0 and 10, the maximum length of the axes for these remaining dimensions \((e)\) is 10, and this value can be used to ensure that non-overlapping hyper-cubes, exhibiting greater similarity within than between each other, are formed around each set of rock type observations (Figure 4.6, middle). The Appendix contains full details of the data preparation.
Interpreting MMD graphs

The MMD technique is used to measure the length of the central tendency of one cluster, or the distance between the central tendencies of two clusters. Calculating the length of a single cluster, or the distance between different clusters, at regular time intervals, and plotting these as graphs of time versus distance, then enables visualization of the clustering trends over time. Clustering trends were calculated for: (1) each geologist’s observed data within each concept, and (2) each geologist’s geospatial position within each concept. Inter-cluster distances were calculated between: (3) pairs of geologists within each concept, and (4) each geologist’s cluster and a prototype cluster for the concept derived from the map legend.

<table>
<thead>
<tr>
<th>Significance</th>
<th>Member</th>
<th>Uncertainty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept Refinement*</td>
<td>- Class Attribution</td>
<td>- Consistent</td>
</tr>
<tr>
<td>*exception: new core due to extreme rise in sampling</td>
<td>- Situation</td>
<td>- Incomplete</td>
</tr>
<tr>
<td>Concept Shift</td>
<td>- Cluster</td>
<td>- Inconsistent</td>
</tr>
<tr>
<td>- Class ATTRIbution</td>
<td>- Situation</td>
<td>- Incomplete</td>
</tr>
<tr>
<td>Concept Stability</td>
<td>- Class Attribution</td>
<td>- Consistent</td>
</tr>
<tr>
<td>- Situation</td>
<td>- Complete</td>
<td></td>
</tr>
<tr>
<td>Concept Stability / Shift</td>
<td>- Class Attribution</td>
<td>- Consistent</td>
</tr>
<tr>
<td>- Situation</td>
<td>- Incomplete</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.7: how to interpret the intra-geologist’s thematic MMD clustering graphs.
Figure 4.7 illustrates how the graphs can be interpreted for (1). A steep downward trend denotes addition of data to the cluster’s central tendency (i.e. core), thus cluster contraction, refinement of a known thematic area, and conceptual consistency; it could also signal conceptual incompleteness as previously unsampled areas of the known cluster are filled out. An exception to this occurs when a new core is formed very distant to the old one, which would actually mean cluster expansion rather than contraction—however, this would require magnitudes of order of new distant samples to be added to the cluster within a time interval, a situation which did not occur in this study.

A steep upward trend signals the addition of novel data to the cluster, sharp cluster expansion, and represents a conceptual shift as well as incompleteness and inconsistency with the prior cluster. A horizontal pattern implies addition of data within the boundary of the cluster, with no particular emphasis on any one part of the cluster, thus signaling consistent and complete understanding of the concept. A gently sloping upward trend indicates mild cluster expansion, thus conceptual consistency but also incompleteness as the cluster continues to add novel data. In terms of pragmatics, cluster expansion indicates knowledge discovery via an intensional pragmatic function and its associated elements (e.g. clusters), while cluster contraction implies knowledge use via an extensional function and associated elements (e.g. class, attribution, situation). However, it is impossible to rule out the presence of the extensional functions during the time intervals exhibiting cluster expansion, simply because it is likely that as part of scientific and cognitive activity the observed data is immediately classified and possibly instantiated. Another caveat is that periods of cluster expansion can help isolate some but not all of the data responsible for concept discovery, as they cannot identify contributing data that exists outside the cluster, such as contrasts. These interpretations can also be applied to (2), with the graphs then showing clustering tendencies of visits to geographically located sites. Interpreting distances between clusters, i.e. for (3) and (4), is more straightforward as upward trends denote cluster divergence, downward trends denote cluster convergence, and horizontal patterns denote stable separation.

4.3.5 Field Data Preparation

Conversion of the field database

Inductive neural approaches such as SOMs are patterned after a neurological signal-response model in which an internal configuration of nodes responds to input channels carrying signals of varying amplitude. The input signals are in this case pre-conceptual and clusters that are representative of concepts are created through stabilization of the internal configuration after repeated exposure to input data. This has proven to be useful for pattern recognition from signal data such as voice recognition, but proves problematic when incoming data is laden with knowledge beyond a simple amplitude measure (Garson, 1998). In such cases the incoming knowledge representation scheme must be explicitly recognized and devolved to a vector of supposedly independent signals for input to a SOM. The devolution of site data to a linear vector is particularly vexing as exploratory field activity may result in a web of relations, where any datum may be related to many others at the same site or between sites. Yet, if conceptual aspects are lost during devolution then any results from subsequent processing must be questioned; hence, adequate and explicit representation of the incoming knowledge is critical to the successful implementation of a SOM. Clearly, if
some concepts are lost or degraded then this also has ramifications for ensuing map accuracy and even for the design of field data collection activities and systems in general.

Significant challenges were posed in converting the field data for input to the SOM, LVQ, and MMD. Although the input data possessed data types such as nominal, interval, ordinal and ratio, more complex arrangements were predominant, such as value domains that were arranged hierarchically (e.g. ‘granite’ ISA ‘igneous’) or were multi-valued (e.g. “biotite, garnet”). These complex attribute domains were themselves grouped into tables possessing intricate relations within the database structure that also required devolution for input to the SOM. See the Appendix for more details.

**Frequency of data occurrence**

An outstanding issue in the preparation of the data involves duplicates in the data. In some cognitive studies, encountering similar data is shown to not affect concept structure. In this view concept structure changes only when the thematic extent of the cluster increases or decreases, not when same data are encountered (Barsalou, 1985). This implies duplicates in the data could be removed, since they will increase clustering and thus skew results, i.e. in both the SOM and MMD, frequency of data occurrence serves to increase clustering toward duplicates. However, scientifically this may not be such a bad thing. Studies of science in psychology have shown that scientists tend to look for confirmatory evidence at different phases of the knowledge cycle (Fiest & Gorman, 1998). The increased evidence then lends weight to the inferences being made and has a role to play in both the discovery and test phases of the cycle: during discovery, frequency of occurrence aids in concept formulation via logical induction, and in the test phase it serves to validate the formulations via pragmatic induction, particularly when disconfirmations are sought. These roles are magnified when the data are sparse and conclusions primarily inferred, as the validations then assume greater significance. Thus, when there are few outcroppings of rock available from which to conceptualize a map unit concept, and individuate its instances, then frequency of observation is important to both concept and individual discovery.

In this chapter it is assumed that frequency of occurrence is an important factor and therefore duplicates in the data are retained. The multiple occurrences represent same observations at different geographic locations. The retention of such occurrences has the added benefit of magnifying certain results in MMD: (1) it serves to highlight concept discovery, in that cluster expansion then signals an important change in the underlying data as the weight of the pre-existing cluster is overcome, and (2) it increases support for theory-driven concepts, as divergence between the cluster and prototype is made more apparent. In future work, multiple occurrences could be removed and the results compared with those in this study.

**4.4 Results**

**4.4.1 SOM: Supervised Classification**

Supervised classification was first performed on six different subsets of the data, as shown in Table 4.1, to test the hypothesis that some of the data is more oriented to concept development than others. The SOM configuration consisted of 500 nodes and 25000 training cycles. The training accuracy was derived using all samples available within the data set, whereas the test accuracy reflects the average clustering results of three random partitions of
the data set into test (220) and learning (Training - 220) samples. The higher overall recognition accuracies of data sets 2-3 indicate these data best fit the map unit concepts.

Table 4.1: recognition accuracy for supervised classification on six subsets of the data.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Training Samples</th>
<th>Training Accuracy</th>
<th>Test Samples</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Dominant Lithology</td>
<td>1535</td>
<td>59.28 %</td>
<td>220</td>
<td>47.42 %</td>
</tr>
<tr>
<td>2 Dominant Lithology and Planar Structural Features</td>
<td>1661</td>
<td>58.52 %</td>
<td>220</td>
<td>55.33 %</td>
</tr>
<tr>
<td>3 Dominant Lithology and Planar Struct. Measurements</td>
<td>1661</td>
<td>65.48 %</td>
<td>250</td>
<td>56.53 %</td>
</tr>
<tr>
<td>4 Dominant Lithology and Structural Measurements</td>
<td>1535</td>
<td>63.44 %</td>
<td>220</td>
<td>49.78 %</td>
</tr>
<tr>
<td>5 Lithology</td>
<td>1466</td>
<td>40.45 %</td>
<td>220</td>
<td>41.36 %</td>
</tr>
<tr>
<td>6 Lithology and Structural Measurements</td>
<td>1466</td>
<td>47.61 %</td>
<td>220</td>
<td>43.78 %</td>
</tr>
</tbody>
</table>

Supervised classification was then performed on data set 3, using different SOM configurations and data subsets, to explore the hypothesis that clustering varied across concepts and geologists. Three subsets of the data were explored: (1) all the data, using four SOM configurations, (2) three partitions of the data according to geologist, using one SOM configuration, and (3) three partitions of the data according to concept, using three SOM configurations. The results in Table 4.2 show that clustering improved within geologists and within concepts, but also that significant overlap is exhibited across concepts. The Sammon view of three concepts, A, B, and C clearly shows this overlap in Figure 4.8.

Table 4.2: supervised SOM classification on all (16) and some (A,B,C) thematic clusters from data set 3.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Samples</th>
<th>Nodes</th>
<th>Cycles</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>1535</td>
<td>300</td>
<td>15000</td>
<td>37.85 %</td>
</tr>
<tr>
<td>ALL</td>
<td>1535</td>
<td>500</td>
<td>25000</td>
<td>55.31 %</td>
</tr>
<tr>
<td>ALL</td>
<td>1535</td>
<td>700</td>
<td>35000</td>
<td>59.35 %</td>
</tr>
<tr>
<td>ALL</td>
<td>1535</td>
<td>900</td>
<td>45000</td>
<td>59.87 %</td>
</tr>
<tr>
<td>ALL-geol 1</td>
<td>563</td>
<td>500</td>
<td>25000</td>
<td>66.61 %</td>
</tr>
<tr>
<td>ALL-geol 2</td>
<td>554</td>
<td>500</td>
<td>25000</td>
<td>55.78 %</td>
</tr>
<tr>
<td>ALL-geol 3</td>
<td>418</td>
<td>500</td>
<td>25000</td>
<td>61.48 %</td>
</tr>
<tr>
<td>A, B, C</td>
<td>959</td>
<td>200</td>
<td>10000</td>
<td>80.60 %</td>
</tr>
<tr>
<td>A, B, C</td>
<td>959</td>
<td>350</td>
<td>17500</td>
<td>79.25 %</td>
</tr>
<tr>
<td>A, B, C</td>
<td>959</td>
<td>500</td>
<td>25000</td>
<td>81.96 %</td>
</tr>
</tbody>
</table>

Figure 4.8: Sammon view (right) of thematic clusters for units A, B and C (black, grey, white).
4.4.2 SOM: Unsupervised Classification

The previous result indicates that geologists’ clusters differ noticeably across all concepts. Unsupervised SOM classification was next performed also on data set 3, to explore whether geologists’ clusters differed within two map unit concepts, A and B.

Figure 4.9: view of thematic spaces for concepts A (left) and B (right). See text for details.

Figure 4.9 illustrates the results in two ways for each concept: as a Sammon mapping that depicts the placement of a geologist’s cluster in the thematic space for a concept, using white, grey and black symbols in the upper left panel; and as a direct view of the SOM array for a concept, where each panel shows a geologist’s data in black contrasted with the remaining two geologists’ data in grey; empty nodes are unused. These remaining panels indicate areas of clear overlap and proximity between black and grey nodes, as well as areas of clear non-overlap and lack of proximity, indicating varying similarity and difference, respectively. Visual inspection also suggests clusters in B are more similar than those in A.

4.4.3 MMD: inter-geologist thematic clustering

The previous results (in 4.4.2) indicate a noticeable level of dissimilarity between three geologists’ clusters for a single concept. To gauge the level of dissimilarity over time, distances between pairs of these geologists’ clusters were measured at weekly intervals, in three concepts, A, B, and C, using the MMD (Mean Median Distance) technique. Each comparison between pairs of geologists represents a line on the graph. The different line lengths indicate that the geologists observed the concept for varying time periods. The results show a general trend to decreased cluster distances, and hence increased cluster similarity, over time for almost every comparison, the exception being a stable pattern for a single comparison in concept C. Despite the trend to increased similarity over time, a consistent and significant level of cluster separation (i.e. dissimilarity) persisted for all pairs (i.e. compare separation distances of > 10,000 in Figure 4.10 with separation distances of 20 between
similar rocktypes in Figure 4.6). The results also corroborate quantitatively the previous results in which geologists’ clusters in B visually appeared to be more similar than A: i.e. clusters in A level out at a separation distance of about 25,000-30,000 while clusters in B level out at a separation distance of about 15,000-20,000.

Figure 4.10: similarity of geologists’ thematic data in three concepts (A, B, C) at weekly intervals.
4.4.4 MMD: intra-geologist thematic clustering

The approximate diameter of a geologist’s cluster was measured at weekly intervals within concepts A, B, and C, using the MMD technique, to gauge the clustering trends of individual geologists. The results in Figure 4.11 show a variety of patterns within and across concepts, for the geologists. For concept B, the geologists’ clusters uniformly decreased over time, indicating an early recognition of the concept, followed by repetitious acquisition of similar data. For concept A, a similar trend to stabilization is interrupted in two geologists by bursts of cluster expansion and contraction, signifying the discovery of new thematic aspects of the concept and intensive re-recognition of the same aspects, respectively. The differences between geologists’ clustering patterns are most pronounced in C, where each geologist displays a different pattern, signifying different levels of uncertainty and completeness.

Figure 4.11: diameter of a geologist’s thematic cluster measured weekly for concepts A, B and C.
4.4.5 MMD: intra-geologist geospatial clustering

Figure 4.12 shows the size of a geologist’s cluster in geographical space. The input data here are not thematic vectors for the observed data, but are instead geospatial coordinates representing the geographical locations for the sites of observation. The graphs in Figure 4.12 show various sampling patterns for the geologists within and between concepts. For concept B, all geologists’ geospatial clusters continuously increase slightly over time, suggesting a slightly expanding concentric pattern of sampling. This is repeated for concept A, but with an upwards swing at one point for each of two geologists, indicating a shift to a new geographical area at that time point. The shifts in location for C are more prevalent and pronounced, as all geologists noticeably change location at one point; for two of them the shift is particularly abrupt and is followed by intense sampling at the new location.

Figure 4.12: diameter of the geologists’ sampling cluster in geographical space.
4.4.6 MMD: inter-geologist prototype clustering

The distance between each geologist’s cluster and the prototypical cluster was measured for each concept. The prototypical cluster was obtained from the text in the legend of the geological map and represents the final description generalized for the concept, as determined and shared by all geologists at the end of the mapping. Performing pair-wise comparison of the prototype to each geologists’ cluster over time, within concept A, B and C, then potentially provides insight into when the prototype is recognized, and whether the concept is in fact data-driven or theory-driven. As illustrated in Figure 4.13, in all cases a significant distance was maintained from the prototype, indicating a radial thematic distribution around the prototype for the concept. Furthermore, two noteworthy observations emerge when comparing the prototype clustering results with the intra-geologist thematic clustering results from section 4.4.4: the two graph patterns are flipped about a horizontal axis for concept B, are identical for C, and varied for A. These patterns are used below to help infer conclusions about the data-driven versus theory-driven nature of the concepts.

Figure 4.13: distance between the prototype and geologists’ clusters for concepts A, B, and C.

4.5 Discussion (I)

These results are interpreted as demonstrating that the situated concepts discovered by the geologists are variably driven by data, theories, and situations, and that geologists exhibit variable individual versions of, and uncertainties about, the concepts.

4.5.1 Data-driven concepts
That some of the geologists’ concepts are partially data-driven is inferred from two indicators: (1) the relative proximity of the geologists’ clusters to the prototype in concept C,
and (2) cluster sizes that are greater than prototype-cluster distances. The first indicator clearly supports this interpretation simply because the prototype and cluster are proximal as one would expect from a data driven concept. The second indicator supports this interpretation because the prototype-cluster distances are consistently shorter than cluster sizes (i.e. compare distances of 8000-13000 for C in Figure 4.13 and cluster sizes of 10000-20000 for C in Figure 4.11), implying that the prototype is located at the core of the cluster, again as one would expect from data-driven concepts where prototypes are central to clusters.

4.5.2 Theory-driven concepts

That some of the concepts are theory-driven is inferred from several indicators: (1) improved clustering as a result of omitting certain types of data; (2) the relatively modest recognition accuracies of supervised classification; (3) the relatively consistent high distance between the prototype and the clusters for A and B; and (4) the obverse patterns between the graphs for the intra-geologist thematic and prototype clustering for B.

The first indicator supports this interpretation because the criteria for eliminating data, and improving clustering, is based on geological theory, which recognizes that certain data is best suited to classification and others for building individual models. In particular, supervised clustering via LVQ improves when significant dimensions are emphasized, such as dominant rock types, and others omitted, such as non-dominant rock types and linear structural measurements; the latter being more suited to characterizing individuals than concepts. Hence theories affect the selection of data to be clustered in the first place.

The second indicator supports the interpretation because data-driven concepts should have higher recognition accuracies, mainly due to good separation of clusters; low recognition accuracies imply a greater overlap of clusters and that the concept is not fully data-driven because theoretical factors are not fully supported by the data.

The third indicator supports the interpretation simply because a consistent larger distance from the prototype implies the prototype is positioned outside the core area of the cluster, as one would expect for theory-driven concepts.

The fourth indicator supports the interpretation because certain reflective patterns also indicate the prototype is not located central to the cluster, i.e. when the overall trend shows cluster size decreasing and distance to the prototype increasing (i.e. compare the inverted trends for B in Figure 4.13 and Figure 4.11), particularly when prototype-cluster distances are greater than cluster sizes (i.e. for B, compare distances of 40000-50000 in Figure 4.13 and cluster sizes of 10000-25000 in Figure 4.11). Note that the inverse does not hold: a pattern in which the cluster size increases and distance to prototype decreases implies centralization of the prototype and thus suggests more of a data-driven concept.

4.5.3 Situation-driven concepts

Response to physical situations and development of individual biases from human situations can be inferred from the results.

Physical situations—heterogeneous environments

Response of at least one cluster to physical situations is inferred from the correspondence in upward shift of the geospatial and thematic clustering, for the same geologist at the same
time interval within the same concept, i.e. in week 6 geologist 3 (solid line) exhibited a significant expansion in cluster size (Figure 4.11) and a significant shift in geographical location (Figure 4.12), while mapping concept C. It is noteworthy that geologist 1 (dotted line) also exhibited a geographical shift in this time interval, and indeed, to the same geographical location as geologist 3 (see Figure 4.14), without a change in cluster size. This reinforces the idea that either geologist 1 had a more complete understanding of C at that time, or was less willing to change.

![Figure 4.14: distribution of data for concept C, by geologist (1=blue, 2=white, 3=yellow); circled area.](image)

denotes observations during weeks 6-7 for geologists 1 and 3.

**Human situations—concept versions**

That geologists developed distinct versions within the same fieldwork perspective for each concept is directly supported by two results: (1) differences in the unsupervised SOM arrays across geologists (4.4.2), and (2) significant distance between geologists’ clusters (4.4.3). Cluster difference itself is not a full indicator of concept variance, inasmuch as the prototypes in theory-driven concepts could overlap while their clusters could differ. Thus, the best evidence for concept versioning is exhibited in C, which is data-driven (i.e. the prototype is central to the cluster) but where the geologists’ clusters do differ. The case for concept versioning is less definitive for concept B, whose clusters are much more distant to the prototype than those in C, and hence are more likely theory-driven. The case for A is variable, as one cluster remains moderately distant to the prototype but two clusters end up reasonably proximal to the prototype, suggesting the differences of the latter two clusters may also be applied to the concept. In summary, there is good evidence to suggest that geologists variably developed distinct versions for the three concepts. There is little evidence to suggest the versions are all due to variable physical situations, hence it might be assumed they result from other factors, e.g. experience, denoted here as human situations.

**4.5.4 Concept Uncertainty**

The intra-geologist thematic clustering graphs for concept C clearly exhibit various uncertainty patterns (Figure 4.11). The horizontal dotted line signifies complete and consistent characterization of the concept. The mildly upward trending dashed line signifies relatively consistent but incomplete characterization of the concept. The jagged solid line
overall signifies incomplete and inconsistent characterization of the concept. The main point being that the geologists exhibited varied uncertainty about C.

4.6 Discussion (II)

Implicit in the techniques deployed above is an underlying suggestion that geoscientific interpretation may be reduced to a computational-quantitative assessment of the supporting field evidence, a claim most likely viewed skeptically by many field-based geoscientists (e.g. Haugerud, 1998). Philosophic and psychological accounts of the scientific process (Feist and Gorman, 1998; Hanson, 1958; Kuhn, 1962; Popper, 1959; Zimmerman, 2000), and considerations of geoscientific interpretation for computation (Dehn et al., 2001; Loudon, 2000; Voisard, 1999), are aligned with such skepticism and with the notion that theoretic and tacit factors extrinsic to data also contribute to interpretation. This resonates with accounts of geoscientific explanation in which situated experience (Baker, 1996; 1999; Frodeman, 1995), interpreted causal processes, as well narrative and diagrammatic descriptions, play a role in defining geoscientific concepts alongside recorded data, and as it were, at times sit above data providing theoretic and experiential superstructure to support them.

The results above corroborate these notions. They suggest several factors affect field-based situated concept discovery: (1) data, as indicated by the partial success of inductive clustering; (2) theory, as evident by the clustering significance of certain attributes and samples and the varying distances to the prototype; (3) physical situations, as response to the environment triggers a conceptual shift in at least one geologist; (4) human situations, or versioning, as suggested by the personal data trends within and between clusters; and (5) uncertainty, as geologists display variable clustering trends. Although these factors are probably not exhaustive, nor is it possible to accurately gauge the amount they uniquely contributed to concept discovery with the methods deployed, it is intriguing and novel to see them revealed from the data and thus empirically supported.

Of particular interest is the way in which these factors relate to our earlier discussion of types of uncertainty, knowledge and conceptualization. Table 4.3 depicts a matrix that is proposed as a framework for these implied relationships: i.e., reliance on theories denotes explicit knowledge and knowledge-driven theoretic conceptualization, as geologists are guided by prior and evolving knowledge expressed as theories. Data-driven influences correspond to probabilistic concepts and also explicit knowledge, inasmuch as summary representations of a cluster that trend to a central tendency are derived from the expressible data. Such theoretic and data-driven influences are clearly connected to, and parameterized by, the uncertainty inherent in explicit scientific knowledge.

Table 4.3: framework for field-based discovery of situated concepts.

<table>
<thead>
<tr>
<th>Theory</th>
<th>Uncertainty</th>
<th>Knowledge</th>
<th>Conceptualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>scientific</td>
<td>explicit</td>
<td>theoretic</td>
</tr>
<tr>
<td>Physical Situations</td>
<td>physical</td>
<td>situated</td>
<td>developed</td>
</tr>
<tr>
<td>Human Versions</td>
<td>human</td>
<td>explicit/implicit</td>
<td>theoretic/?</td>
</tr>
</tbody>
</table>
The development of ingrained personal versions, or *individual bias*, is clearly connected to the subjective aspects of human uncertainty and thus is most likely characterized by both implicit and explicit knowledge. The *explicit* component of individual bias may be related to personal theories held, and hence to theoretic conceptualization. The *implicit* component of individual bias may refer to some combination of: (1) formerly explicit knowledge that has become ingrained and automatic after repeated usage (e.g. rock type identification), and to (2) tacit, perhaps holistic impressions generated from exposure to situations. Although there is no concept development process clearly attributed to such implicit knowledge, investigations into perception and implicit learning hint at several possibilities (Barsalou, 1999; Goldstone, 1998; Goschke, 1997). Also, the overt individual biases we observed are particularly significant in light of the efforts undertaken by the individuals to ensure compatibility of methodology, terminology, concepts and clusters during mapping. Their presence underscores the capacity for mappers to develop different conceptual or individual models for the same map area, particularly when compatibility between their models is not reinforced. Finally, the impact of the physical environment on concept development relates to situated knowledge, physical uncertainty and to developed clusters.

4.6.1 GeoScientific Computing

The preceding discussion suggests that geoscientists implement various types of knowledge and concept discovery mechanisms when interpreting field data, and that the resulting conceptual models, individual models, and maps are constrained by various uncertainties inherent in the data or reflected by the data. A requirement for geocomputational systems then is the accommodation of the broad range of knowledge activities involved in scientific explanation, including the facilitation of various modes of representation and reasoning. Specifically, the computational issues that arise from these conclusions are manifold, but they basically involve: (1) developing enhanced mechanisms for representing geoscientific conceptual and individual models, implicit, explicit and situated knowledge, and human, physical, and scientific uncertainty; (2) developing mechanisms for geoscientific concept discovery and evolution that incorporate probabilistic, theoretic and developed conceptualization functions in the course of reasoning; and (3) combining the representation and reasoning methods into a unified framework for geoscientific computing.

The concerns about representation as well as reasoning raised in this study are poignant: while some implicit and situated knowledge may be difficult, if not impossible, to express and represent, the remainder of such knowledge should be optimally structured for storage in databases and for use in geocomputational reasoning. Furthermore, detecting the missing knowledge components and understanding how, for instance, they affect map contents and accuracy is also critical for interoperating with geoscientific databases, an increasingly important issue.

4.7 Conclusions

Field data from three individuals were analyzed using geocomputational techniques. The results suggest that concept development in the field involves data, theory, situations, versions and uncertainty. These results lend general support to claims in previous chapters about the impact of pragmatics on conceptualization, by demonstrating that the process of
concept discovery and evolution is impacted by methodologic and epistemic factors. In particular, the results corroborate proposed notions about the existence and nature of situated concepts (i.e. that they exist and are situation-driven) and the existence and nature of concept versions (i.e. that they exist and differ for epistemic reasons, such as personal viewpoints). In this way they also respond to the third geo-pragmatic challenge, and to the third objective, both of which call for empirical substantiation of geo-pragmatics.

The results also inform chapter 5 by providing a means by which some of the data leading to concept discovery can be uncovered, and thus providing a means to populate the schema designed in that chapter with the origins and effects of the concepts studied in this chapter. The results also indirectly support chapter 6, which proposes situated concepts occupy a prominent role in the structure of ontologies. In general the results corroborate some assumptions and theory laid out in previous chapters, and provide empirical support, analytical mechanisms and data utilized in future chapters.

4.7.1 Potential benefits related to the geo-pragmatics of mapping

An unexpected result of this work is the novel deployment of the MMD technique and associated metrics, used here to evaluate another relatively novel aspect of map quality—the quality of the situated concepts themselves as understood by a group of mappers. This notion of evaluating map concept quality is fundamentally geo-pragmatic, in that conventional approaches treat such concepts as essentially static, complete and well-defined, without need of evaluation, whereas this study indicates that in fact they can be dynamic, partial, situated and uncertain—that they essentially can be parameterized by the discovery process. Then having at hand some methods and tools to evaluate this type of quality becomes important to both the field-based construction of maps, and the appropriate use of such maps.

Potential benefits to mapping include the ability to select team members prior to mapping, based on prior clustering patterns in certain physical situations, as well as the ability to monitor the divergence and convergence of individual understandings during mapping, particularly when individuals are not co-located. Indeed, use of the methods and metrics may reduce the need for extensive co-location and allow mapping teams to cover more ground when there is evidence of convergence, to re-convene upon conceptual divergence, and even to pinpoint at which times and places such discrepancies began in order to achieve resolution. Note these teams might be robotic instruments, operating, say, on Mars... There are also training implications as the methods might be used to pinpoint areas in which training (or repair) are required.

Potential benefits to map use include the move toward a quantitative basis for determining the reliability of a map, or map feature, a problem initially highlighted in chapter 1 as fundamentally geo-pragmatic. One could image developing reliability metrics related to the uncertainty, coherence and completeness revealed by the clustering patterns. Of course, these would not be prescriptive, and could not be used in isolation, but would certainly improve on current mechanisms that are ad-hoc, subjective and qualitative.

4.7.2 Future Work

These results also represent a first-order empirical glimpse at the knowledge building process in geological fieldwork, and they set the stage for more detailed and focused studies—this study is in effect a reconnaissance. The results also indicate several points of emphasis for
the development of geocomputational systems for field and other geoscientific information, highlight the utility of geocomputational tools for their study and processing, and raise significant questions about the reliability and uncertainty inherent in geoscientific maps. In general, future work should address the expansion of geocomputational frameworks to accommodate a greater variety of reasoning, representation and visualization techniques that can be brought to bear not only to geoscientific data but also to other space-time information (Gahegan, 2000b).

More specifically, further work could also investigate how the prototypes themselves are developed, and compare their development to cluster development, to better differentiate theory-driven tendencies in concept development from data-driven tendencies. Combining this with in situ studies that shadow mappers, or follow-up interviews, would provide better controls on the interpretation of the clustering results. This would not only potentially help with insights into prototype and cluster development, but also possibly into the impact of frequency of occurrence in scientific studies, and the transitions between concept development (clustering, blending) and instance development (instantiation, individuation, classification).
CHAPTER 5: STRUCTURING GIS DATABASES FOR GEO-PRAGMATICS

This chapter bridges the results of chapter 3 and chapter 4 by applying the formalism developed in chapter 3 to the empirical data studied in chapter 4. In particular, a conceptual schema is developed for representing the theoretical geo-pragmatic elements. The general intended application of the conceptual schema is database design, though it could hypothetically be adapted for other systems such as ontologies. The schema directly models many aspects of both the origin and effects of geoscience concepts and instances, resulting in these elements being directly included in the database contents. This represents an abstraction level above the level typical for database representation. This shift is due the fact that in many geoscientific disciplines concepts are regularly discovered, evolved and versioned, but the architecture of geospatial information systems is primarily aimed at supporting static conceptual structures in which concepts are predetermined and do not change in step with scientific progress. The result is a gap, addressed in this chapter, between our evolving understanding of these concepts and how they are represented into our systems.

The schema is represented as a UML diagram (Universal Modeling Language; Rumbaugh, et al., 1999) and its prototype implementation by both the author and third parties (e.g., Ballesteros, 2004; Soller et al., 2002) is summarized in this chapter. The author’s prototype implementation uses a common desktop GIS (i.e. ESRI’s ArcView). It also uses the data from the experiments described in chapter 4 to populate the schema, and utilizes other results from the same chapter to discriminate between the origins and effects of the geological unit concepts studied therein. These are then visualized in geographic space using the desktop GIS, to illustrate the pragmatic aspects of the concepts and to demonstrate the resultant enhancements to their representation. In doing so, this chapter addresses the fourth objective of the dissertation, in that it illustrates one way that the pragmatic aspects can affect geospatial database design. Likewise, it also addresses the second significant geo-pragmatic challenge identified in chapter 2 concerned with the impact of geo-pragmatics on schematics, by exploring some of the ramifications of the abstraction shift in geospatial database design.

5.1 Introduction

The importance of determining what geographic concepts exist and how to represent and process them computationally is a significant research thrust in GIScience. This thrust is leading to richer and more complete conceptual models in our systems, but is also requiring us to extend the representations we employ for capturing knowledge about our domains and recording our understanding of them. In particular, advanced representations must tackle the thorny issue of missing knowledge, in that much of the knowledge required to validly interpret information stored in GIS, and indeed in other information systems, by both humans and machines is implied and not explicit (Rubenstein-Montano, 2000)—it depends on various tacit agreed conventions to enable the communication of meaning between information producers and consumers. Common carriers of meaning are the names and definitions we give to concepts, but these capture only a fragment of their meaning. Our representations must therefore become richer to reduce misunderstandings between producers and consumers. On the one hand this involves explicitly representing more of the meaning residing with producers, such as representing the pragmatic aspects of entities, and on the other hand, it involves communicating or reconstructing this meaning in information consumers from such explicit representations.
Database systems are particularly susceptible to the meaning gap, because databases are primarily intended to manage information, not concepts, and they consequently assume that all concepts are predetermined and static. As discussed previously, this is problematic in the geosciences where situated concepts are regularly discovered, evolved, and versioned, and hence treated as information. Indeed, maintaining databases of situated concepts, such as for geological formations, ecologic units, soil units, etc., is a significant responsibility of various government agencies. One approach to overcoming this problem in database systems is to treat situated concepts as a form of information recorded in the database (e.g. Mennis, 2003; Mennis & Peuquet, 2003). In this approach, enhanced meaning could be further achieved if a concept’s pragmatic aspects are also recorded alongside it in the database, in addition to the concept itself, and this requires development of a schema for modeling concepts and their pragmatic aspects, such as the one proposed in this chapter. Furthermore, although such a schema might be directed primarily at databases, it might in fact apply to any ontology in which situated concepts are represented.

This chapter is organized as follows: section 5.2 proposes a metamodel approach to the representation of geographic concepts; section 5.3 develops a UML schema for this approach based on the geo-pragmatic elements identified in chapter 3; section 5.4 implements the schema using the information and results from chapter 4; section 5.5 describes partial implementations of the metamodel approach by third parties, and the chapter concludes in section 5.6.

5.2 Geoscientific Databases

A developmental and situated viewpoint on geoscientific ontology has significant consequences for geoscientific database design, in that database schema founded on evolving and versioned concepts will themselves need to change, leading to maintenance and usage headaches (Roddick, 1995). The solution proposed here involves database schema founded on meta-concepts, such as ontologies, models, concepts, etc., rather than on unstable concepts in the domain. A database schema developed on this principle will have the benefit of being a repository for dynamic (situated) concepts and a registry for multiple scientific knowledge components and their relations, and could include an enhanced pragmatic representation of concepts. This presupposes that at least some geoscientific knowledge components, such as situated concepts, are dynamic and contextualized, or that a meta-concept organization for database schema is preferable for reasons of being able to deal more effectively with change, or to convey a deeper understanding of where aspects of meaning originate. A dynamic and contextual account of geoscientific concepts involves:

- discovery/evolution/versioning: a static view of geoscientific knowledge does not accommodate the learning or discovery of new knowledge, or facilitate the multiple interpretation or re-interpretation of existing data—fundamental objectives of any science. Modeling an open system, such as geoscience knowledge, with the closed world assumptions is inherently problematic (Frank, 1997); embedding such assumptions in fixed database schema will inevitably lead to a program of continual schema adjustments, imprecision and heterogeneity, or to pre-set limits on the types of knowledge acceptable, escalating representational inaccuracy and hindering scientific creativity.

- generality/specificity: concepts and theories are thought to range in generality, from widely applicable to domain specific (Guarino, 1997; Rosch, 1978). This range of
generality suggests that in the lower conceptual tiers concept change might be more prevalent. Conceptual change at the highest levels may be non-existent, whereas at intermediate levels it may be infrequent and regarded as a paradigm shift (Kuhn, 1962), and instability in less general concepts may be seen as ongoing human learning, such as the evolution of scientific understanding of some region, or as natural evolution of the region over time leading to shifts in the associated situated concepts. For example, the general concept of ‘lake’ may remain relatively fixed within a perspective, but the concept of ‘polluted lake’ may be socially situated and evolutionary, altering with changing definitions of pollution, with evolving lists of pollutants and developments in pollution measurement, and with fluctuating pollution levels linked to environmental conditions.

The premise that conceptual instability entails schema change arises from the practice of founding schema on concepts extracted from the domain at a specific level of abstraction. Of the five levels of abstraction identified by Brachman (1979), three are relevant here: the epistemological level contains concept structuring rules and primitives such as tuples, relations, objects, classes, attributes, slots, etc.; the conceptual level contains concepts identified in the domain, their properties, relations, and constraints; and the linguistic level contains data and relations. These levels apply to database design in a top-down fashion: at the first level epistemological frameworks are selected; at the second level concepts are elicited from the domain and represented using an epistemological framework in three ways: as: (1) a technology-neutral conceptual schema, (2) a technology-aware logical schema, and finally, as (3) a technology-specific physical schema intended for a particular hardware/software system; the third level is the level of data, which resides in the physical system. This narrowing of technology refers to increased commitments made at each level to a specific hardware and software environment, ending with one system at the physical level. Thus, schema developed upon unstable concepts identified at the conceptual level will inevitably be prone to change, as seems to be the case with many scientific databases (Tamzalit & Oussalah, 1999; 2000).

In contrast to versioning mechanisms that focus on managing schema change (Roddick, 1995), this chapter concentrates on a broadly applicable conceptual schema design founded on general concepts that are presumably more stable. Such concepts might be drawn from an additional level, the ontologic level (Guarino, 1994; 1995), which serves to increase the meaning of epistemological or conceptual elements by connecting them to broader conceptual-logical systems, or top, domain/task and application ontologies, ranging in generality from universal to increasingly specific. However, though general ontologic concepts might be identified within a geoscientific domain, the nature of open systems, the knowledge discovery imperative, and situated concepts, quite on their own and without reference to domain characteristics, argue for the need to model unexpected and variable domain relations and properties that cannot be fully predefined, or when approximated, result in a complex network-like schema structure that is difficult to use and maintain. In effect, it is difficult to ascribe global regularities of structure to domain objects and relations in open systems, such as to situated concepts that are subject to discovery and change.

To overcome these limitations the primitives of the more abstract epistemological level, called meta-concepts, are used as the basis for schema design. Specifically, a technology-neutral UML conceptual schema for concept and data interaction is developed, one that
might be logically and physically adapted in databases or in other applications that represent concepts, such as geoscience ontology systems.

As part of raising the abstraction level for database schema, both the top-down ontologic approach from concepts to individuals, and the bottom-up situated approach from individuals to concepts, is incorporated, as modeled by the clustering, classification and instantiation functions in the pragmatic square (of section 3.2). This bidirectional relationship contrasts with other unidirectional, top-down, geospatial approaches that introduce spatial and/or temporal constructs at the epistemological or ontologic levels (Benslimane, et al., 2000; Camara, et al., 1994; DeOliveira, 1997; Fonseca et al., 2000; Hadzilacos & Tryfona, 1996; Kosters, et al., 1997; Pullar & Stock, 1999; Renolen, 2000; Shekar, et al., 1997; Smith et al., 1991). It also contrasts with non-geographical meta-representations in which the link is mainly unidirectional, either top-down (e.g. Gruber, 1993; Noy et al., 2000; Pepper, 2000; Tudhope et al., 2001), or bottom-up but not pragmatic (e.g. Fayyad, 1996; Brachman et al., 1999; Wille, 1996).

The inclusion of such bottom-up factors is beneficial in many respects. As described later in section 5.5, modeling situated concepts in the schema improves database design by providing a more normalized structure and by leading to greater semantic precision in the representation. Apart from these design benefits, the schema also results in significant use benefits, as it enables scientific discovery to be represented and hence tracked in computing environments. For example, applying the schema to the case study from the previous chapter, as was done here, enables queries on the database to quickly and easily display and contrast origins and effects of concepts, in this case for groups of geologists or single geologists. Visualization of these aspects, either in terms of their geospatial location (as was done here) or their thematic location (as in the previous chapter), could then inform map accuracy and quality by showing where geologists converged or diverged in their understanding of concepts, and where understanding is complete and consistent within the group or within single mappers. This could then aid measuring the effectiveness of field teams, selecting experts for specific tasks, etc.

5.3 Concept Structure

Recall that general interest in concepts has foundations in two main traditions, the philosophical and cognitive, which respectively emphasize logical and mental representations of concepts. In both traditions concepts possess intension and extension: extension refers to the group of entities considered to exemplify the concept, whereas intension typically refers to the essential meaning encapsulated by the concept usually expressed as properties, constraints, definitions, etc. In both traditions intension and extension are the main components of concept structure, and this general approach can be characterized as being largely semantic, inasmuch as how a concept originated, was used, or what it affected is not included (for exceptions in cognitive science see Barsalou, 1999; Smith & Samuelson, 1997). However, these latter considerations are pragmatic, and as argued in previous chapters, are required for augmented meaning representation. For example, an intension in the pragmatic representation described in section 3.3 refers not only to essential characteristics but also to their origins. Some common approaches to semantic representation are reviewed and schematized in UML in section 5.3.1 below, and the schema is embellished with pragmatic aspects in section 5.3.2.
5.3.1 Semantic concept structure

Figure 5.1 depicts a fragment of a schema that shows the traditional relationships between a concept and its extension. Note that there is agreement between the cognitive and philosophic traditions on aspects of the meaning of these relations: “where there are distinct kinds of categories, the associated concepts will also be distinct” (Medin, et al., 2000), indicating a particular extension applies to a single concept, and that concepts need not have extension but may be abstract e.g. ‘quality’. Moreover, individuals can be classified in multiple ways as “it is logically possible for one and the same list of objects to be in one-to-one correspondence with the extensions of substantially distinct concepts” (Sutcliffe, 1993).

Note that a concept’s properties are not modeled in Figure 5.1, as that is beyond the immediate purpose; it is recognized that a concept’s properties might be structured in many ways, perhaps using frames, objects, slots, conditions, etc. (see Barsalou & Hale, 1993). Also, implied, but not shown in Figure 5.1, are semantic relations between concepts, and individuals. The semantic relations include parthood, dependence, spatial, temporal, etc., but critically they do not include pragmatic considerations such origins for any element. This puts the basic elements in place for what will be called here “conceptual networks”, which are networks of concepts, individuals and their semantic relations. Examples of such conceptual networks include various traditional schema, such as database schema, document schema (e.g. XML), etc., and models containing instances from such schema such as databases, documents, etc.

<table>
<thead>
<tr>
<th>Concept</th>
<th>intension</th>
<th>extension</th>
<th>Class</th>
<th>instance</th>
<th>Individual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1..1</td>
<td>0..1</td>
<td></td>
<td>0..*</td>
<td>classification/instantiation</td>
<td>1..*</td>
</tr>
</tbody>
</table>

Figure 5.1: the traditional relations between a concept and its extension.

5.3.2 Pragmatic concept structure

A pragmatic approach to concept structure implies that a concept’s origins, effects and uses should be included in its structure. As discussed in chapter 3, the constituents of a pragmatic representation include pragmatic functions and their inputs and outputs. Table 5.1 repeats Table 3.2 here to summarize the interaction between these constituents. Table 5.2 in addition summarizes their definitions. Understanding their interactions then provides insight into the external relations for a concept, which informs schema design.
Table 5.1: pragmatic functions and reasoning in the pragmatic square (repeated from Table 3.2).

<table>
<thead>
<tr>
<th>Pragmatic Mapping</th>
<th>Pragmatic Function</th>
<th>Source Entity Type</th>
<th>Target Entity Type</th>
<th>Entity Collection for sources, except $F', G'$</th>
<th>Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F, G$</td>
<td>clustering</td>
<td>individual</td>
<td>concept</td>
<td>cluster</td>
<td>induction</td>
</tr>
<tr>
<td>$f$</td>
<td>blending</td>
<td>concept</td>
<td>concept</td>
<td>blend</td>
<td>abduction</td>
</tr>
<tr>
<td>$g$</td>
<td>individuation</td>
<td>individual</td>
<td>individual</td>
<td>situation</td>
<td>abduction</td>
</tr>
<tr>
<td>$F', G'$</td>
<td>instantiation</td>
<td>concept</td>
<td>individual</td>
<td>attribution</td>
<td>deduction</td>
</tr>
<tr>
<td>$F'^{-1}, G'^{-1}$</td>
<td>classification</td>
<td>individual</td>
<td>concept</td>
<td>class</td>
<td>deduction</td>
</tr>
</tbody>
</table>

Table 5.2: summary of the key constituents of geo-pragmatics.

<table>
<thead>
<tr>
<th>constituent</th>
<th>summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>concept</td>
<td>scientific artefact that is an abstraction over one or more entities</td>
</tr>
<tr>
<td>individual</td>
<td>single entity that retains its identity over time</td>
</tr>
<tr>
<td>state</td>
<td>description of an individual at some time interval</td>
</tr>
<tr>
<td>cluster</td>
<td>collection of individuals used to develop a new concept via clustering</td>
</tr>
<tr>
<td>blend</td>
<td>collection of concepts used to develop a new concept via blending</td>
</tr>
<tr>
<td>class</td>
<td>collection of individuals that are classified as exemplifying a concept</td>
</tr>
<tr>
<td>instance</td>
<td>member of a class, placed in the class via classification</td>
</tr>
<tr>
<td>situation</td>
<td>collection of individuals used to identify a new individual via individuation</td>
</tr>
<tr>
<td>attribution</td>
<td>collection of individuals whose properties were modified via instantiation</td>
</tr>
<tr>
<td>intension</td>
<td>relation between a concept and its origin consisting of a cluster or blend</td>
</tr>
<tr>
<td>extension</td>
<td>relation between a concept and its class (an effect)</td>
</tr>
<tr>
<td>situated</td>
<td>relation between an individual and its situation (its origin)</td>
</tr>
<tr>
<td>instantiated</td>
<td>relation between a concept and its attribution</td>
</tr>
<tr>
<td>clusterMembership</td>
<td>relation between a cluster and its members (individuals)</td>
</tr>
<tr>
<td>blendMembership</td>
<td>relation between a blend and its members (concepts)</td>
</tr>
<tr>
<td>situationMembership</td>
<td>relation between a situation and its members (individuals)</td>
</tr>
<tr>
<td>attributionMembership</td>
<td>relation between an attribution and its members (individuals)</td>
</tr>
<tr>
<td>instanceOf</td>
<td>relation between a class and its members (individuals)</td>
</tr>
<tr>
<td>classification</td>
<td>deductive function for placement of an individual into a class</td>
</tr>
<tr>
<td>instantiation</td>
<td>deductive function for transfer of properties from a concept to an individual</td>
</tr>
</tbody>
</table>
individuation: abductive function for identifying an individual via differentiation from others
clustering: inductive function for development of a concept from individuals
blending: abductive function for development of a concept from other concepts
perspective: ontology developed using a specific methodology within a domain
version: variation of an entity within a perspective due to epistemic factors
context: recursive collection of origins, effects and uses for an entity
interpretation: context with a single path recursion

### 5.3.3 Conceptual schema design for geo-pragmatics

The schema shown below in Figure 5.2 and Figure 5.3 includes all pragmatic elements listed in Table 5.2 except for contexts and interpretations, which are dynamically constructed, e.g. as queries in database systems. The concept and individual elements in the schema are shorthand for the general case of abstractions and their instances, respectively. This shorthand is used because the geological mapping case study used in the implementation below mainly requires two levels of abstraction, situated concepts and individuals, though in actual fact the schema accommodates multiple levels of entity abstraction. To accommodate the general case, “concept” could be renamed to e.g. “abstraction”, and “individual” to e.g. “worldEntity”, to denote an abstraction and its instances, and to reflect the fact that a concept or individual could serve in either role, as abstraction or instance; this becomes important as additional types of abstractions are identified in chapter 6. States, on the other hand, are at the lowest level of abstraction and can only be instances of higher levels. The current schema naming is simply convenient for the case study and prototype implementation below. Importantly, when this schema is applied and populated with data, the resulting structures are called here “pragmatic networks” as they detail the origins, effects and uses of the concepts in addition to the elements modeled by conventional conceptual networks. Pragmatic networks are thus a superset of conceptual networks in which pragmatic elements are added to conceptual networks. Apart from this structural distinction, there is also a temporal distinction: pragmatic networks are essentially historical, documenting how an element came to be, what effects were generated as a result, etc., in contrast to conceptual networks which are basically ahistorical and document how elements can be related at any time.

![Figure 5.2: conceptual schema for geoscience pragmatics, Part A.](image-url)
5.4 Implementation

The conceptual schema shown above in Figure 5.3 and Figure 5.2 is implemented for the geological mapping case study described in chapter 4. The results of this prior chapter are used to identify origins and effects for the situated concepts discovered in the case study, and these are then represented and visualized in map form with a GIS. Specifically, ArcView GIS is used for representing and visualizing the geometry and topology of the geospatial properties of individuals, and the bulk of the schema is represented in the MS-Access relational database and linked to ArcView via the identifier for individuals. These systems were chosen mainly because they were conveniently available.

Although the case study implicitly contains multiple levels of abstraction (i.e. individuals, situated concepts, theoretical concepts), the implementation considers only those entities explicitly expressed on the map, i.e. situated concepts and individuals, where individuals are conflated with states. Also, only one perspective applies, a fieldwork perspective, though three versions are considered for both concepts and individuals, one for each geologist whose interpretations formed the subject of the experiments in the previous chapter. Because of these considerations, the actual implementation in MS-Access is somewhat simplified, with perspectives and versions mapping to attributes of concepts and individuals, rather than to stand-alone tables. Furthermore, the instantiation function and related entities such as attribution and attributionMember are not included in the MS-Access implementation, because this information could not be derived readily from the case study. The physical
implementation in MS-Access is shown in Figure 5.4 below. Some tables, e.g. for the pragmatic functions, are not shown in order to simplify the diagram.

Figure 5.4: physical implementation in MS-Access of the conceptual geo-pragmatic schema.

Results of the previous chapter are used to inform and populate the schema, which in turn is used to geographically visualize and corroborate some of the results from the previous chapter. Specifically, the time-indexed thematic clustering graphs in chapter 4 are inspected for: (1) areas of cluster expansion, which imply discovery of property concepts or clustering in this case, and (2) areas of cluster stability or contraction, which imply solely extensional development such as individuation and instantiation.

However, recall some limitations to this technique:

- Periods of classification cannot be isolated, as classification can apply to all graph patterns. This is because it is likely that concept development immediately triggers classification in fieldwork, as part of basic cognitive and scientific function, hence it is unreasonable to differentiate clustering from classification within the weekly time interval of the case study.

- Cluster membership is incomplete because relevant contrasts are not identified by this technique. Hence clusters mainly consist of class instances in this implementation, with the cluster forming a subset of the class of property concepts—i.e. only some of the property observations contribute to clustering, but all property observations are classified.

Table 5.3 shows the results of this inspection, which is used to populate the schema.
The following sections then geo-visually present aspects of the populated schema, to demonstrate its viability. Only geo-pragmatic elements such as class and cluster, related to a single ontologic concept, ‘Unit C’, are presented. This involves not only the ontologic concept ‘Unit C’, but also its conflation with the concept in the fieldwork perspective, i.e. ‘Unit Cfw’. Also included are three version concepts, each generated by a distinct geologist, i.e. ‘Unit Cfw1’, ‘Unit Cfw2’, ‘Unit Cfw3’, and a concept for the Unit’s properties, i.e. ‘Unit Cfw-properties’, and its three versions, i.e. ‘Unit Cfw1-properties’, and so forth. Recall that in the geological mapping case study instances of the ‘Unit Cfw’ concept are visually presented as polygons on the map, and that various observations—instances of properties of ‘Unit Cfw’—are presented as points.

5.4.1 Class—Effects of ‘Unit C’

Figure 5.5 and Figure 5.6 illustrate the classes for the concepts and properties of ‘Unit Cfw’, respectively, and for their versions as developed by the three geologists.

Figure 5.5 shows the geographical extents of the instances of the cumulative ‘Unit Cfw’ concept, in yellow, as well as its three versions. The versions corroborate the analysis in chapter 4, visually showing that significant overlap occurred between the classes, i.e. most but not all of the polygons were mapped by each geologist. Figure 5.6 shows the geographical position of the instances of the property concepts for each version of ‘Unit Cfw’. The instances denote observations of rock type and its geometrical orientation, and the concepts denote generalizations of those properties as assigned to the unit concept. The geo-visualization shows a fairly uniform geospatial distribution of sites of observation for the three geologists.

An outstanding question focuses on the degree of individuation and instantiation at those property instances that are in the class but not in the cluster—i.e. given that the properties of the unit concept are established, does the pattern of property observation then change to characterizing the individual rather than the concept, i.e. mapping the geographical boundary of the instance and describing it, rather than learning the thematic boundaries of the concept?
Figure 5.5: from top left, clockwise, classes (in yellow) for: ‘Unit Cfw’, ‘Unit Cfw1’, ‘Unit Cfw2’, ‘Unit Cfw3’.

Figure 5.6: classes for versions of properties of ‘Unit Cfw’: ‘Unit Cfw1-properties’ (blue), ‘Unit Cfw2-properties’ (white), ‘Unit Cfw3-properties’ (yellow).
5.4.2 Cluster

The cluster for each version of the property concept for ‘Unit \( C_{fw} \)’ is presented in Figure 5.7. These clusters represent the observations taken during the time intervals associated with property concept discovery, as revealed by the thematic clustering graphs in chapter 4. Note that for the cumulative property concept, ‘Unit \( C_{fw}\)-properties’, the geospatial distribution is broad and the number of observations limited, indicating that property concept discovery occurs over few samples taken from the breadth of geographical locations, as one would expect from situated concepts. However, also note that as revealed in chapter 4, not all geologists needed to sample each area before determining the concept.

![Figure 5.7: clusters for versions of properties of ‘Unit \( C_{fw} \)’: ‘Unit \( C_{fw1}\)-properties’ (blue), ‘Unit \( C_{fw2}\)-properties’ (white), ‘Unit \( C_{fw3}\)-properties’ (yellow).](image)

5.4.3 Context, interpretation, blend, and situation

A context and interpretation for ‘Unit C’, the blend for ‘Unit \( C_{fw} \)’, and a situation for instance 1008 of ‘Unit \( C_{fw} \)’ are shown in Figure 5.8, Figure 5.9, and Figure 5.10. Figure 5.8 shows these geo-pragmatic elements as a pragmatic network. The complete network is a context for ‘Unit C’. The solid open arrows signify a blendMembership relation and the dashed closed arrows signify both an instanceOf and clusterMembership relation because the entities at the tail of dashed arrows are members of both the class and cluster, respectively. Then, the entities at the head of the arrows denote both a class and cluster, and serve as a proxy for the concept. An interpretation, blend and situation associated with ‘Unit C’ are then shown as members of the context, i.e. the members of an interpretation for ‘Unit C’ are enclosed by the
dashed line, the members of the blend for ‘Unit Cfw’ are enclosed by the dotted line, and some of the members for the situation for individual 1008 of ‘Unit Cfw’ are enclosed by the solid line. Figure 5.9 then illustrates how the blend is stored in the database, and Figure 5.10 geo-visually presents the situation, including two versions, one per geologist. Note that the context and situation as depicted in Figure 5.8 are incomplete, because the properties for one of the two versions are not depicted—i.e. the properties associated with the first geologist, ‘Unit Cfw1-properties’, are omitted for reasons of space. However, the full suite of property observations for the situation is shown in Figure 5.10, including instances and assumed contrasts, with the latter referring to sites adjacent to the polygon instance and hence presumably used to infer its boundary.
5.5 Other implementations by third parties

Apart from the limited prototyping described above, the main implementers of aspects of this schema are other researchers striving to improve representation of geological maps. These parties are mainly interested in structuring geoscience databases using logical and physical variations of the conceptual schema detailed above. In all cases the schema is adopted only partially, and in most cases the origins and instantiation aspects are not deployed. In terms of content, the implementers are primarily interested in modeling situated concepts and their classes, hence in the metamodel approach generally advocated in this chapter. The one exception to this is a research project at ITC (International Institute for Geo-Information Science and Earth Observation, the Netherlands) that extends an earlier version of the schema to store inferences in the development of situated geoscience concepts, but without identifying the types of inferences or functions as done here (e.g. classification, etc.)

All parties are experiencing the same benefit, namely improved spatial database design. The original problem collectively facing them was their adoption of a cartographic viewpoint to database design. In this viewpoint, situated concepts are located in the legend of a map and are not described in the database. Hence they are seen as being useful only for specifying symbolization rules and labels for the individuals that appear on the map. This exhibits a lack of recognition that the individuals on the map are actually instances of the situated concepts described in the legend. The focus on individuals for the contents of the databases is also due
in part to the database systems deployed, which are engineered mainly to describe individuals with geometries, e.g. polygonal geological regions and point-based observations, and do not readily accommodate the representation of related concepts.

The cartographic viewpoint and emphasis on individuals led to both database design and scientific issues. Central to these issues was the fact that one description template for an individual, containing one set of properties, was forced to accommodate two descriptions, one for the individual and another for the related situated concept—sometimes simultaneously. In practice, the description for an individual rarely consisted of information unique to the individual; more often than not it was populated by information from the situated concept, which was prototypical for all individuals that are instances of the concept; and in some cases both were awkwardly shoe-horned into one description. Hence, this led to: (1) semantic confusion in which it was impossible to determine whether the individual description was unique, prototypical, or both; (2) data incompleteness, as not all geoscience knowledge was available in the database, when a prototypical or unique description is omitted; (3) normalization issues, in which awkward data structures (in 1st normal form) were used, e.g. when situated concept descriptions were repeated for many individuals, thus wasting space and causing problems on updating when the concept evolved. By teasing out the distinction between concepts and individuals in the database schema design, the third parties are now able to avoid such problems, and build databases with higher semantic precision and that follow classical database engineering principles. This in turn allows them to store a broader suite of data (i.e. both individuals and prototypes) and to avoid semantic confusion in applications that rely on the data. The latter is particularly important due to the fact that geological maps are quickly becoming an important layer in various decision-making processes; then providing the appropriate level of detail for the application can be crucial. Semantic confusion is further discussed in chapter 6.

The third parties include members from the geoscience academic community, and industry, as well significant federal geoscience data providers in the U.S. and Canada. These parties extend an earlier version of the conceptual schema intended for application to geological map information (Brodaric & Hastings, 2002). Variations of this map-oriented schema are now being implemented in: the Canadian Knowledge Network (CGKN; www.cgkn.net), USGS National Geologic Map Database (NGMDB; Soller et al., 2002), the Georgia Basin Digital Library (GBDL; Talwar et al., 2003), the ESRI geologic map data model and the research project mentioned above at ITC. In each case the schema forms the basis for a conceptual database design that is modified for logical and physical designs for the project. The implementations focus on different challenges: CGKN emphasizes the concept representation challenges by populating the schema with both situated and theoretical geoscientific concepts; NGMDB and ESRI focus on the map representation challenge, using the schema to drive map-based visualizations of geological information; GBDL focuses on the plurality challenge by representing the multiple perspectives on sustainable development in some communities; and the research project at ITC concentrates on capturing pragmatic aspects, namely inferences made during field mapping. Each project is briefly described below.

### 5.5.1 An object model for geologic map information

Figure 5.11 depicts a schema for geologic map information (Brodaric & Hastings, 2002) influenced by two sources: (1) an earlier version of the schema described above (Brodaric & Gahegan, 2002), and (2) an effort to create a standard geological map data model for North
America (Johnson, et al., 1999; NADM, 2004). The resultant schema is geared to representing information associated with a geological map from the viewpoint that a map denotes a visual presentation of a geoscientific model, one that may utilize several existing ontologies and be derived from prototypical evidence gathered in the field and elsewhere.

As the details are relayed elsewhere (Soller et al., 2002), only its main elements are summarized below. The core components include symbols, concepts, occurrences (individuals) and descriptions. Note this schema adds the following items to the geo-pragmatic schema: properties (descriptions), cartographic markers (symbols), and models for collections of concepts, symbols, occurrences (individuals), and descriptions; it does however lack the origin and effect aspects:

- **concept**: refers to scientific abstractions.
- **occurrence (individual)**: refers to individuals conflated with states.
- **symbol**: refers to the elements comprising the cartographic display. Both concepts and individuals can be symbolized. E.g. mines can be designated with a default symbol that can be overridden for a specific mine occurrence. Symbols may be assigned cartographic rules for scale dependencies, etc., via specific software.
- **description**: numeric, textual and possibly spatial properties of a concept or individual. E.g. a textual or geometric description of a map unit, the numeric and/or textual details and location of a specific measurement or observation.
- **symbolic (cartographic) model**: a collection of symbols representing a cartographic palette, such as a symbol standard for a mapping agency.
- **conceptual model**: a collection of concepts, their relations and descriptions corresponding to a specific vocabulary, taxonomy, classification scheme, or ontology.
- **occurrence model**: a collection of classified and described individuals.
- **description model**: a collection of property descriptions, often corresponding to a data set.

The collections, or models, enable the definition of a legend as a symbolized conceptual model, and the definition of a map/geological model as a complex model composed of a legend and aggregations of individuals. A geological model might then be defined by omitting the symbolic components. These definitions and resulting schema can serve as a knowledge layer over information. It also has the potential to be serve as an ontology repository that can facilitate information integration: e.g. a conceptual model can represent a mixed axiomatized-prototypical ontology, a specific description model can represent a linguistic ontology, an occurrence model can represent the instances being integrated, and their various relations represent the mappings between ontologies.
Figure 5.11: conceptual schema for geologic maps (adapted from Brodaric & Hastings, 2002).
5.5.2 The Canadian Geoscience Knowledge Network

CGKN is a cooperative initiative of the Canadian government geoscience community that aims to provide a national web-based portal to its information holdings. CGKN is being constructed as a loosely-coupled network of heterogeneous databases distributed across the country. Interoperability in two sub-projects, the national bedrock and surficial map databases, is being developed through the use of a common schema, derived from the schema above (Davenport et al., 2002; Struik et al., 2002).

A major challenge faced by both sub-projects is the acquisition from existing maps of national-scale, situated map unit concepts and their descriptions. The major issues involve managing versioning and lexical ambiguity: map unit concepts often have different descriptions (versions) in various geographic regions, individual descriptions vary according to the mapper, and different descriptive terms often have identical meanings while identical terms often have different senses. The versioning problem is being handled via concept relations, similar to the approach advocated above. Solutions to the lexical problem are progressing in both top-down and bottom-up modes: top-down normative definitions are being developed for many of the terms and at the same time, in a bottom-up approach, existing terms are being input into databases with a view to finding empirical regularities in the use of the terms and coordinating these with the definitions. The aim is to retain local terms and characteristics, as these apply directly to the maps, but to fit these terms into a uniform system for interoperability purposes, with the view that such a system would grow and evolve with scientific advances and shifts. Consequently, the situated nature of the information itself, and the expected dynamic nature of the geoscientific ontology being envisioned, requires a system that can provide contexts to facilitate the understanding of concepts and eventually the map objects.

The schema fragment most heavily utilized by CGKN is that for storing and describing situated concepts and individuals, which often share descriptive properties such as those for age and rock type. The sub-projects have collected and entered into the schema several thousand map unit concepts and descriptions, as indicated in Table 5.4.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Sub-Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geological Province</td>
<td>2</td>
</tr>
<tr>
<td>Tectonic Terrane</td>
<td>17</td>
</tr>
<tr>
<td>Tectonic Assemblage</td>
<td>42</td>
</tr>
<tr>
<td>“Regional Unit”</td>
<td>99</td>
</tr>
<tr>
<td>(Supergroup, Supersuite)</td>
<td>2</td>
</tr>
<tr>
<td>“Regional subunit”</td>
<td>159</td>
</tr>
<tr>
<td>Group, Suite, Complex</td>
<td>43</td>
</tr>
<tr>
<td>Formation, Lithodeme (+ informal equivalents)</td>
<td>161</td>
</tr>
<tr>
<td>Member (+ informal equivalents)</td>
<td>21</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2082</strong></td>
</tr>
</tbody>
</table>
5.5.3 The USGS National Geological Map Database

The NGMDB project is congressionally mandated to develop a database of geological map information for the U.S., in cooperation with state and other agencies. NGMDB is being developed in three stages:

- metadata: a web-based index of published maps, stratigraphic names and current mapping projects (see http://ngmdb.usgs.gov/);
- spatial: a web-based catalog of raster and vector map images; and
- non-spatial: a database of geological content underlying the spatial data.

The schema shown in Figure 5.11 above was adapted and prototyped in NGMDB’s third stage (Soller et al., 2002). The prototype emphasized using the schema as a framework for visualizing geological information in map form. Implementation in an object-oriented GIS (GE-SmallWorld™) demonstrated the schema is an effective means of hosting and visualizing multiple geological maps (models) from a single information source. Advanced display operations were added to objects to enable sophisticated visualization: e.g. MapView→Display displays a particular geological model (i.e. concepts, individuals, etc.) from the information source; MapView→ReClass re-classifies a map according to a property (e.g. rock type, as in Figure 5.12); and Occurrence→Display performs scale-sensitive display of individuals: e.g. teeth appear on faults at more detailed scales, and certain map units are grouped together at regional scales through generalization of the unit to a more general unit.

![Figure 5.12: reclassifying map units into rock types (from Brodaric & Hastings, 2002).](image)

5.5.4 The Georgia Basin Digital Library

The GBDL aims to foster awareness, understanding and dialog about sustainability amongst municipal and scientific communities in the Georgia Basin, a region in southwestern British Columbia, Canada (Harrap et al., in press; Talwar et al., 2003). GBDL provides a web-based mechanism to host and browse natural resource and socio-economic information intended for input into modeling tools that generate future urban scenarios. Such scenarios are useful to regional planners and policy-makers during decision-making. GBDL uses a variation of the schema described above to organize 64 natural resource geospatial data sets (i.e. spatial models), sustainability concepts and multi-media descriptions (i.e. textual, image-based). It
emphasizes multiple sustainability ontologies, including those reflecting community, academic and NGO perspectives. Figure 5.13 displays the ‘Ideas & Perspectives’ module where the sustainability ontologies can be explored by browsing a conceptual network of relevant concepts (top left) whose selection triggers the display of related maps (bottom left) and multi-media descriptions (right).

The map-oriented schema, populated with sustainability concepts and their descriptions, figures prominently in the overall technological architecture of GBDL as illustrated in Figure 5.14. The architecture possesses three tiers: (1) the information tier implements the schema physically in a relational database system (SQL ServerTM); (2) the services tier provides a web-service to access to the sustainable development ontologies and descriptions, and it provides standard geospatial web-services for accessing geospatial individuals; and (3) the presentation tier provides a user interface to the library (information tier) and its functions (services tier), exclusively using the services tier to access the information.
5.5.5 ESRI Gelogic Map Data Model

An experimental physical implementation of the schema was developed by the ESRI GIS company (Grise & Brodaric, 2004). The main driver for this implementation was the idea of re-using a metamodel and situated ontologies not only for geology but also in other geoscientific domains. A fragment of the simplified physical model is shown in Figure 5.15, and experiments in navigating information presented in this model are shown in Figure 5.16. The latter, visualization and navigation remain, an open challenge for such designs.

![Figure 5.14: the three tier architecture of GBDL.](image)

![Figure 5.15: ESRI geology data model fragment (from Grise & Brodaric, 2004).](image)
Figure 5.16: visualizing an occurrence (left) and its concept (right) in the experimental ESRI data model (Grise & Brodaric, 2004).

Figure 5.17: logical instantiation of the map-oriented conceptual model (from Ballesteros, 2004).
5.5.6 Metamodel instantiation for geoscientific data collection

The original ideas behind both the map-oriented data model and the geo-pragmatic schema are tested in an M.Sc. thesis (Ballesteros, 2004). The emphasis is placed on modeling the inference process between concepts and individuals, in order to represent their origins. Figure 5.17 depicts the resultant logical data model. This logical model is being adapted to field data collection systems in ongoing research at ITC.

5.6 Conclusions

This chapter examines the schematic implications of geo-pragmatics. It develops a conceptual schema to model the geo-pragmatic entities identified in chapter 3, and implements the schema in a GIS and relational database. The information and results from chapter 4 are used to populate the schema, and some geo-pragmatic elements are visualized in map form for demonstration purposes. Additional implementations of an earlier version of the schema by third parties are also described. The focus in these implementations is the inclusion of situated geoscience concepts and individuals directly in the database, in order to handle their discovery, evolution and versioning. The chapter addresses the fourth objective, and second geo-pragmatic challenge, by exploring the impact of geo-pragmatics on geoscience schema.

The main points introduced by this chapter are:

- the schematic representation of situated concepts benefits from a metamodel approach, which models these concepts directly in the schema, because such concepts are prone to discovery and change due to methodological and epistemic factors;
- the schema can be implemented in common GIS environments, and such implementations improve the explanation of geoscientific entities via the map-based visualization of geo-pragmatic elements;
- aspects of the metamodel approach are being widely tested by third parties;
- testing in other domains and with more data is required to determine if, as assumed, the approach is transferable and scalable.
- benefits of implementing the schema include the ability to quickly view origins and effects of concepts, which can aid in the evaluation of the accuracy and quality of maps and mappers. This can help not only deciding which map or mapper to trust for which type of concept, for specific tasks, but it can also inform ongoing collaborations to help guide and coordinate them. These benefits are primarily founded on: (1) the ability to make semi-quantitative visual evaluations for issues of map quality that were previously implicit or largely qualitative, and (2) by applying visualization, graphing and statistical tools to data stored in the schema.

This overall metamodel approach is potentially applicable in many domains where knowledge is regularly discovered, evolved, and versioned. It could therefore be used for other applications such as the dynamic visual exploration of concepts and associated source information. Computability of the recursive context structure is an ongoing concern. Future work entails increased testing with diverse information, generating better data capture tools, and implementing within emerging ontological systems.
CHAPTER 6: LEVELS OF ABSTRACTION IN GEO-PRAGMATIC ONTOLOGIES

The last two chapters focused on the empirical and schematic aspects of geo-pragmatics. They demonstrated concept versioning in situated concept discovery, proposed some pragmatic functions for generating such versions, and suggested a schema that incorporates the functions to explain the versioned concepts. This chapter can be seen from a very practical viewpoint as an exploration of how to organize the concepts that could be stored in that schema. In doing so, and unlike the previous two chapters that focus on the human pragmatics of concepts, this chapter emphasizes natural pragmatics as a means of organizing concepts, i.e. the role of foundational factors in reality that lead to and explain a variety of concepts. Because such factors are a priori such exploration is intrinsically ontologic. What is proposed in this chapter then is a foundational framework, called LEKXIS, for organizing geographical concepts into six levels of abstraction. Each level is suited a specific type of concept, which is defined, as are the conditions for differentiating concept instances. Importantly, one of these levels contains situated concepts, which are dependent on their origins in geospace-time, and hence on their geo-pragmatics. Situated concepts represent the conceptual aspect of situational knowledge, which is introduced in chapter 2 and chapter 3. This then helps define theoretical knowledge more precisely as those concept types occurring at levels that are more abstract than the situational. It also extends existing notions of ontology structure and highlights the role of the geographic via the inclusion of situated concepts.

The roots of LEKXIS are discussed from a variety of perspectives, and the resultant framework is presented and demonstrated using examples from geoscience. Ontology implementation is demonstrated via representation in two popular ontology syntaxes: UML and OWL. This chapter responds to the fifth objective of the dissertation (from chapter 1), and also to the first major geo-pragmatic challenge (from chapter 2), which both deal with the impact of geo-pragmatics on semantics.

6.1 Introduction

Information ontology (henceforth “ontology”) contents are loosely organized into three levels of abstraction in previous chapters: states, individuals and concepts. In this chapter the concept level is further differentiated into four levels, resulting in six levels in total. This is distinct from existing approaches to ontology organization that only recognize three or four levels of abstraction. For example, some well-known approaches to ontology structure use philosophical and computational principles to help identify general entities from which specializations can be derived in all domains, and rules and guidelines for relating the specialized entities within a domain (e.g. Gangemi, et al., 2003; Grenon & Smith, 2004). This leads to a 3-tier vertical ontology structure consisting of: (1) the upper-level for general entities, (2) the domain level for entities specialized within a domain from some upper level entity, and (3) individuals, with the latter typically considered as a token that is typed by entities from the other levels, e.g. ‘substance’ ‘Mountain’ ‘Mt. Whistler’.

Sometimes a fourth level is inserted between domain entities and individuals to reflect, ostensibly, concepts that do not apply across the domain but which may be local to specific applications such as particular tasks, usages or information requirements (Guarino, 1998) e.g. ‘substance’ ‘mountain’ ‘MyConceptOfMountain_forLandslideAssessment’ ‘Mt.’
Whistler'. Because of its loose definition this level might conceivably even be stretched to also accommodate situated concepts constrained by specific geospace-time situations and shared histories amongst instances, e.g. ‘substance’ ↔ ‘mountain’ ↔ ‘Rocky_mountain’ ↔ ‘Mt._Whistle’, or ‘substance’↔‘species’↔‘human’↔‘Boyan’. Then, it can be interpreted as being pragmatic because of its apparent situation dependence: it appears to be driven by functional, causal or geospace-time context. However, such an interpretation would extend the level beyond its original sense, which would seem to be motivated more by computational concerns that lack firm ontologic grounding. This lack of ontologic grounding may be the reason why a situated ontological level has been largely overlooked in the more recent computational ontology designs rooted in philosophical rather than strictly computational principles.

What is proposed in this chapter is a refined ontology organization that consists of six levels for the vertical structure of ontologies, including one that provides ontologic grounding for a level that is situated, i.e. context sensitive, pragmatic and geographically oriented. The envisioned benefits include increased semantic precision and granularity. Semantic precision increases because conditions are added for categorizing concepts into the levels. Semantic granularity increases because the levels are ordered in increasing levels of abstraction. These outcomes should facilitate ontology design by providing additional levels for the hierarchical organization of ontologies, and consequently should improve geoscience knowledge capture as the levels correspond to abstractions made implicitly by geoscientists. In addition, the inclusion of a level grounded in geographical context would also underscore the fundamental contribution of the geographical sciences to ontology design as a whole. A use-case demonstrating the benefits of increased semantic granularity is discussed, but the remaining benefits will remain implied and subject to future work.

The intent of the chapter is to present the conceptualization and some examples using existing concept representation languages and geoscientific use cases. No attempt will be made to formalize the conceptualization as a logic system; this is also left to future work.

6.1.1 The need for increased granularity in vertical geoscience ontology structure

Geoscience ontologies are gaining prominence not only because of their role in information interoperability, but also for their role in cyberinfrastructure in which hypotheses are created and tested, where ontologies form parts of various theories (e.g. plate tectonics), classification systems (rock classification), and so forth. The role of such theoretical knowledge in e-geoscience is especially significant because it is not simply an end product, as in classical sciences such as physics and chemistry, but is also a means to infer models about the natural history of an area from observed or measured samples. However, because of gaps in both samples and knowledge and due to various sampling methodologies, it is possible to infer multiple valid models for an area (Schumm, 1991). Theoretical knowledge then not only serves to help generate models, but also to select optimal models based on the fit of data to theory.

Concepts are a component of such knowledge, and are considered here to be scientific abstractions that vary in terms of generality. Examples of geoscience concepts that span much of this range of generality include ‘unit’ or ‘material’, ‘formation’ or ’granite’, ‘formation X’ or ‘granite of X’, ‘this rock body instance of X’ or ‘granites of this rock body instance of X’. These concepts abstract either across geographic situations or within specific
situations. This proves problematic for existing ontology organizations that do not distinguish this range, nor provide clear mechanisms to vertically organize it. The result is a disparity in the semantic granularity and overall quality of the ontologies being constructed, and a disparity between the ontologies and the semantic granularity evident in the domain, which ultimately hinders ontology use because of this disconnect. As an example of the latter, consider the models used by geoscientists in practice might be grouped here into theories, conceptual models and physical models. Theories apply across the domain and reside in textbooks, classification systems, and professional standards (e.g. plate tectonics, rock type classification systems). Conceptual models denote a specific geospace-time situation in terms of local concepts and their geospace-time relations, but without specifying any geometries or positions in geographic space. Physical models, on the other hand, reconstruct the state of physical objects in geospace-time, including their geometries, positions and changes in time, without necessarily describing their generalized concepts (e.g. a 3D model or simulation typically describe only individuals and states). Importantly, these distinctions are not simply lexical, as the different geoscience models represent distinct entities actively used in unique ways by geologists. Capturing existing geoscience knowledge, or building effective interfaces to enable use of such knowledge in cyber-infrastructure, then requires mappings between the practical conventions of the domain and an adequately grained ontology structure. But this ontological granularity does not exist in current ontology design practices, where all domain concepts are either lumped into a single level, as introduced above, or are spread across an unlimited number of levels, as discussed below. If the practical uses of geoscientists are truly to be reflected and enhanced for e-geoscience, then it is necessary to disambiguate these levels of abstraction to enable effective knowledge capture and use.

The desire for greater granularity in vertical ontology structure is motivated not only by geoscience, but also by other application domains in which complex vertical abstraction structure is evident but not fully explained, leading to unclear implementation guidelines for ontology organization. Examples of this include the biological domain, e.g. ‘species’ → ‘human’ → ‘Chris’ (Guarino & Welty, 2002), or the manufacturing domain, e.g. ‘car model’ → ‘Mustang’ → ‘this Mustang’ (Sowa, 2000, pp. 32), where domain ontology contents are commonly separated into vertical levels of abstraction by instanceof relations, but where the reasons for doing so are unclear or difficult to use. One approach to explain these levels relies on meta-properties, i.e. identity, unity, rigidity, and dependence, to guide the instantiation and subsumption of a concept by another (Guarino & Welty, 2002: ‘human’ is discerned to be an instance of ‘species’ due to differing identity conditions, i.e. instances of ‘human’ are identical if they co-locate in geospace-time, whereas instances of ‘species’ are identical if they co-locate in a biological classification; so ‘species’ cannot subsume ‘human’, because subsumption implies same identity conditions, and ‘human’ must be an instance of ‘species’. A geoscience analog is ‘formation’, ‘formation X’ and ‘this rock body instance of X’, where instances of ‘formation’ (e.g. ‘formation X’) are identical if they share the same prototypical type section, and instances of ‘formation X’ (e.g. ‘this rock body instance of X’) are identical if they co-locate in a geospace-time region.

While these meta-properties ably aid the hierarchical placement of concepts within ontologies, particularly for the subsumption (ISA) relation, there are certain drawbacks. First, identity conditions are hard to find. Second, finding them requires knowing in the first place what a concept’s instances are: e.g. finding an identity condition for C requires knowing that A and B are its instances prior to analyzing how to differentiate them; conversely, if A and B
represent a class of entities differentiated by an identity condition, but we do not know which concept instantiates A and B, then its not clear where to assign the identity condition. Third, the number of levels remains undefined, allowing for an indefinite number of levels, though in practice few are used (Sowa, 2000).

These drawbacks make it harder to re-apply these patterns of organization in other domains, such as natural science mapping, where similar structures seem to hold, e.g. map unit rankings such as geological ‘formation’, soil ‘series’, or ecologic ‘domain’ seem analogous to ‘species’; map unit classes such ‘formation X’, ‘Polar Domain’ (Sorokine, et al., 2004) and ‘Dennis Series’ seem analogous to ‘human’, and individual objects on maps such as ‘this rock body instance of X’, ‘this instance of the Polar Domain’ and ‘this instance of the Dennis Series’ seems analogous to ‘Chris’. Thus, while significant progress has been made in explicating aspects of vertical organization associated with the subsumption relation (e.g. ‘human’ ISA ‘mammal’; Guarino & Welty, 2002; Bittner et. al., 2004), the lack of explanation regarding other aspects (e.g. ‘species’< ‘human’ or ‘formation’< ‘formation X’) leaves open questions about the number of abstraction levels and the principles for placing an entity at a level.

In this chapter, necessary and sufficient conditions are provided for each level, to enable classification of a concept into a level, e.g. to place ‘formation’ and ‘formation X’ at the appropriate level. Identity conditions are also provided for each level to enable differentiation of entities at the next lower level. The levels may be also considered meta-concepts, because they circumscribe the nature of the concepts for a level, and the guiding conditions may be considered meta-constraints for the meta-concept.

The chapter proceeds as follows: section 6.2 summarizes existing approaches to vertical ontology structure and identifies outstanding issues; section 6.3 introduces a richer suite of concepts and describes the resultant abstraction levels, referred to as LEKXIS (pronounced as “lexis”), as driven by [L]ogic, sch[E]ma, content [K], space-time conte[X]t, [I]ndividuals and [S]tates; Sections 6.3 also provides an example of LEKXIS, and discusses related issues; section 6.4 examines the organization of LEKXIS in terms of vertical and horizontal relations in ontologies; section 6.5 discusses potential applications, and section 6.6 concludes with closing comments.

6.2 Levels of abstraction—existing approaches

The focus on ontology structure follows a historical shift in AI, depicted in Figure 6.1, from processing considerations to concerns about the nature of the entities being processed. This shift is most famously espoused by Newell (1982) who identifies the knowledge level in addition to hardware and software levels, to account for knowledge content, and independently by Brachman (1979) who subdivides knowledge content into five levels, including the conceptual level for concepts of interest, and the epistemological level for concept structuring primitives such as classes and properties. According to Guarino (1995) the ontology level occupies an additional stratum between the conceptual and epistemological levels and is intended to constrain the organization of concepts using philosophically founded invariant meta-properties and entity types. For example, three kinds / levels of ontologies are introduced by Guarino (1998): upper-level ontologies apply to all domains, domain and task ontologies apply within a single a domain, and application ontologies apply within certain activities or processes within a single domain.
For the purposes of this chapter, domain ontologies can be seen as organized into an orthogonal grid consisting of vertical and horizontal levels. Horizontal levels represent width-wise segmentation of the grid into disjoint thematic partitions possessing unique identity criteria (e.g. ‘physical’, ‘biological’, ’social’; Guarino, 1999), whereas vertical levels represent height-wise segmentation into abstraction partitions that are typically linked via instanceOf relations (here denoted by the ‘→’ symbol, e.g. ‘species’→’mammal’ in a horizontal biological partition). Internal organization of a specific cell in the grid involves the ISA relation for vertical structuring (e.g. ‘human’ ISA ‘mammal’) and other relations for horizontal structuring such as parthood, constitutes, co-located, and so forth.

### 6.2.1 Horizontal levels of ontology

Horizontal levels of ontology are thought to consist of disjoint partitions for material, mental and social entity types (Bateman, 1995; Poli, 2001), each of which can be further subdivided: e.g. material entities into properties and objects (Frank, 2003), or physical, biologic, etc. (Guarino, 1999), and these can be further stratified to account for spatial-temporal co-location (Donnelly & Smith, 2003). Mental entities may be subdivided into symbolic, conceptual and perceptual levels (Gardenfors, 2000), and finally, social entities encompass linguistic elements, including those revealed via ethnographic studies (Mark & Turk, 2003). These horizontal levels of ontology are separated by gray boundaries in Figure 6.1.

---

**Figure 6.1:** horizontal (solid) and vertical (dashed) levels of ontology.
6.2.2 Vertical levels of ontology

Horizontals levels of ontology suggest nested disjoint thematic partitions in which vertical levels of abstraction may apply. However, vertical organization within the material, mental, and social partitions have little in common superficially and indeed exhibit conflicting underlying principles, i.e. material entities such as universals (correlates in the environment) and particulars (individuals in the environment) are mind-independent and grounded in natural and social necessity; mental entities (concepts, conceptions) are grounded in mental processes, and linguistic entities (terms, propositions, sorts) are grounded in language constructs and use. Furthermore, each possesses disjoint levels as depicted in Figure 6.1. In some accounts material entities may be vertically organized into particulars, universals and formal categories (Gangemi, et al., 2003; Grenon & Smith, 2004), conceptual entities into spaces, concept regions and object points (Gardenfors, 2000), and linguistic entities into lexicogrammar, semantics, context (Bateman, 1995).

To understand the broad similarities and differences between these perspectives, and to help identify issues, specific work within each perspective is further analyzed below. The analysis considers how three core components are related. The components consist of abstractions, objects, and representations, where abstractions denote e.g. types, concepts, categories, etc., objects denote source entities from which abstractions are derived and/or to which they refer, and representations are the means by which abstractions and objects are represented.

Material approach to vertical ontology structure

The material approach recognizes three entity types, and three corresponding abstraction levels: particulars are individuals in the real world, universals are meaningful patterns found in individuals, and formal categories are logical abstractions of universals (Figure 6.2).

![Figure 6.2: three abstraction levels implied by material ontology: particulars, universals, categories.](image)

Particulars and universals are vertically connected via a primitive instanceOf relation, and an extension is defined as the group of particulars thus linked to a universal. Universals and categories are (presumably) hierarchically connected via the ISA relation (Bittner et al., 2004; Neuhas et al., 2004), defined intensionally as the relation between a universal or category and their abstract partitions, which reflect the extensional relations between an extension and some sub-group of it that is also an extension (Bittner & Smith, 2003). These entity types and levels roughly map onto kinds of ontologies, inasmuch categories constitute upper or formal
ontologies and universals constitute domain or material ontologies (Grenon & Smith, 2004). Computer or mental representations of such ontologies can then be considered variably faithful descriptions of underlying reality (Smith, 2004). However, one shortcoming of this design for present purposes is the coarseness of the vertical structure of universals, which does not distinguish between cases such as ‘human’ as vertically related to ‘species’ and ‘human’ as vertically related to ‘mammal’, neither in terms of type of vertical relation (both are presumably ISA) nor type of entity (all are generic universals).

**Logical language approach to vertical ontology structure**

Unlike the material approach, the logical language approach possesses no pre-ordained entity types that determine vertical ontological structure. It instead relies on the nature of the vertical relations (ISA, instanceOf) to generate a structure from given elements in a domain. As is well known, the vertical relations are defined relative to the definition of a class. A class is represented symbolically by a unary predicate, denoted by a set of individuals in one or more worlds (actual, possible, intended) that are collectively its extension, and connoted by an intension which does not necessarily alter when the extension changes. The vertical relations of subclass (ISA) and instanceOf are then defined in terms of being a subset or member of the extension, respectively (Figure 6.3). Extension members are called class instances and may even be classes themselves (e.g. ‘human’→’species’ in languages such as KIF, OWL-FULL and others). However, there is little constraint on what qualifies as an instance, with the criteria seeming to devolve to the specific needs of an application. As a result, there is no limit placed on the length of instanceOf chains, i.e. sequences of classes connected via instanceOf, even though common practice suggests otherwise. Operational guidelines recommend at this point that such chains consist of a single instanceOf relation, and hence two vertical levels, for purposes of soundness, completeness and decidability of reasoning (e.g. OWL-DL), but this does not specify what content should be included at a given level. In short, due to lack of constraints on the instanceOf relation the number of vertical ontology levels is undetermined a priori, and the grounds for initiating a level are arbitrary, offering minimal guidance to ontology construction in this respect. And without such guidance, it is likely that different groups of ontology developers will construct ontologies using different (or ad-hoc) guiding principles.

![Diagram](image)

**Figure 6.3:** indefinite abstraction levels via logical languages.

130
Cognitive approach to vertical ontology structure

In the cognitive approach, ontology contents and structures reflect mental entities and structures, respectively. Mental entities are derived from embodied interaction with the world, vertical levels are determined via mental abstraction over such entities, and language meaning is closely tied to the resultant structures (as in cognitive linguistics; e.g., Langacker, 1999). The primary carriers of meaning in mental structures are concepts, variably determined in different approaches by: (1) necessary and sufficient conditions, (2) their role in theories, (3) prototypical or actual examples of world entities, (4) dynamic constructions in situations, or (5) inherent mental atoms (Barsalou, 1999; Komatsu, 1992; Laurence & Margolis, 1999; Murphy, 2002).

An important theory of conceptual structure, based on a view of concepts as prototypes, is the notion of conceptual spaces (Gardenfors, 2000). Vertical organization in conceptual spaces involves three entities and three corresponding levels: spaces, concepts and objects (Figure 6.4). Conceptual spaces consist of dimensions often possessing metrics and are analogous to frames (Minsky, 1975), albeit with topological and geometric structure; concepts (including properties) are regions in a conceptual space and are often defined in terms of proximity to a prototypical object; and objects are points in a space. The ISA relation can then be defined as a subspace relation between concepts (Ahlqvist, 2004). The instanceOf relation is undefined, though can be hypothetically viewed here as the relation connecting levels. However, as it stands for ontology design, there are no clear grounds to help inform what to place at a level, e.g. in ‘biological rank’⇐’species’⇐’human’⇐’Chris’ which entity denotes a space, concept, or object, and why? Furthermore, what is not easily accounted for is variation in conceptual structure between levels, such as the instantiation of properties from a higher to lower level, e.g. ‘shape’ might include ‘number of sides’ as an essential dimension, but ‘triangle’ needs dimensions for ‘side 1’, ‘side 2’ and ‘side 3’ (Funkhouser, forthcoming).

An alternative approach to knowledge representation that does provide partial explanation for abstraction is founded on semiotics (Bateman, 1995; Sowa, 2000). In this approach, semiosis refers to the process of abstraction in which object-abstraction-representation triads are generated from other such triads within some interpreter, e.g. <Chris, concept-of-Chris (i.e.
‘Chris’), “Chris”> might give rise to <Human, concept-of-Human (i.e. ‘Human’), “Human”> (Figure 6.5). Semiosis is thought to apply widely, including to geoscientific abstraction (Baker, 1999), but for purposes of ontology design the link between triads, hence between abstractions, is underspecified and could be interpreted to be ISA, instanceOf, or possibly something else, and the number of variations and thus potential levels of abstraction is open and indefinite. The same holds for cognitive theories in which abstractions are derived from mental simulations of previously experienced situations (e.g. Barsalou, 2003).

Figure 6.5: a semiotic view of abstraction in which O-A-R triads are derived from other triads.

Summary of existing approaches to vertical ontology structure
The starting point for existing approaches to vertical ontology structure is either declarative or functional. Declarative approaches assert a fixed set of three primitive entity types and associated levels, whereas functional approaches claim a single primitive abstraction process. In all cases where described, the ISA relation applies within an abstraction level and the instanceOf relation connects levels. The primitives, levels and connecting relations differ between approaches, with minor overlap and with different advantages and shortcomings. Of particular significance is the fact that none of the approaches satisfactorily account for the abstraction chain ‘biological rank’ ‘species’ ‘human’ ‘Chris’. Therefore, the approaches also do not each fully account for similarly structured chains such as those in the geological domain, e.g. ‘geological unit rank’ ‘formation’ ‘formation X’ ‘a rock body of X rock’ or ‘mountain’ ‘fault-origin mountain’ ‘a Rocky Mountain’ ‘Mt. Whistler’. A basic requirement for improved vertical ontology structure then is the development of a framework that explicates these chains in order to inform ontology content and its vertical organization. Such a framework should notionally fuse pertinent aspects of declarative and functional approaches, maximizing benefits and minimizing shortcomings. The LEKXIS framework is an initial attempt to satisfy these requirements. Its declarative aspects are discussed in the next section, including identification of primitive meta-concepts and associated abstraction levels, while its functional aspects are largely undetermined at this time.
6.3 Levels of abstraction in LEKXIS

The discussion above accentuates differences in vertical ontology structure between representation approaches grounded in the material, mental and social perspectives. However, other viewpoints in these perspectives imply greater similarity across perspectives and suggest loose, but not exact, convergence to a common set of core entity types and grounding conditions. Integration of these additional viewpoints is for the most part absent from computer-based ontology representations. So, as a first step to such representation, recall that it is proposed here that this convergence, derived from analogies amongst the viewpoints, involves meta-concepts roughly grounded in logical conditions [L], schema [E], content or properties [K], context [X], individuals [I], states of individuals [S], and that these distinctions form the basis for ontology levels. Key elements of the additional viewpoints that lend support to this analysis are sketched next.

Material perspective revisited
In her treatise on substances, Millikan (2000) identifies the following substance types:

- **Substances** [L]: are formal categories valid across domains. They represent knowledge about subjects: natural artefacts, human artefacts (concrete or social), matter, organisms, and indeed anything that can be re-identified, bear properties, and about which useful inferences can be made;
- **Templates** [E]: are schemas that frame entities;
- **Eternal kinds** [K]: are ahistorical entities with essential properties or causes;
- **Historical kinds** [X]: are products of evolution or causation in a geospace-time region;
- **Individuals** [I]: are singly realized subjects in the environment that might have time-slice states [S]. Individuals and states are not explicitly included in Millikan’s ontology, but are mentioned as being pertinent.

Cognitive perspective revisited
Reviews of mental concepts typically include the concept types listed below (e.g. Komatsu, 1992; Laurence & Margolis, 1999; Murphy, 2002). Because of the varying underlying principles and cognitive mechanisms typically associated with each of these types, they have been generally considered disjoint and incompatible (for one exception see Barsalou, 1999). Of importance here is that some evidence has been found to support each of these types, though none totally accounts for aspects of how humans form concepts and classify entities.

- **Classical** [L]: categories are grounded in necessary and sufficient conditions;
- **Theoretical** [E]: categories are grounded in (1) the role they play in theoretical structures, or (2) in their schematic structure (e.g. image schemata; Lakoff, 1987) possibly leading to upper-level cognitive invariants (e.g. ‘on’);
- **Prototype/exemplar** [K]: categories are grounded in the graded similarity of properties to a core representation;
- **Situated** [X]: categories are grounded in dynamic constructions made in the situation; these are heavily influenced by prior experience (Barsalou, 1999; Smith & Samuelson, 1997).
Linguistic and semiotic perspectives revisited

Linguist and semiotic analysis often, and non-exclusively, identifies the following levels in the study of the meaning of some text:

- syntactics [L]: the syntax, grammar and perhaps lexicon of a language, often defined axiomatically;
- schematics [E]: position within a text fragment, the structure of a document, or the structure of a logical proposition representing the text fragment;
- semantics [K]: link to a referent in a model or world, or to a linguistic type or sortal;
- pragmatics [X]: linguistic context and language use.

6.3.1 LEKXIS

Table 6.1 suggests levels of ontology for the types of entities identified as substances by (Millikan, 2000), which is similar to the notion of substances in (Grenon & Smith, 2004). LEKXIS consists of levels in which: [L] has formal concepts, [E] has schematic spaces, [K] has partitions of schema spaces that are atemporal, acausal and non-geospatial (ahistorical) (after Gardenfors, 2000, and Millikan, 2000), [X] has situated concepts that are historical, i.e. historically-causally tied to a situation in a geospace-time region, [I] has entities with persistent identity in a geospace-time region, and [S] has time-slices of individuals. These levels lead to a vertical organization of concepts within a domain, and the concepts within a level are examples of the meta-concept that defines the level. The names for the meta-concepts are used interchangeably here to refer to both the vertical level of an ontology in which concepts are grouped, as well as to the meta-concept itself.

<table>
<thead>
<tr>
<th>Ontology Kind</th>
<th>Ontology Entity</th>
<th>Ontology Perspective</th>
<th>LEKXIS Level</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper Formal</td>
<td>category</td>
<td>substance, classical, syntactic</td>
<td>Upper (Logic) [L]</td>
<td>formal categories grounded in logical or possibly cognitive invariants.</td>
<td>- substance - material artefact - social artefact - physical quality - physical object</td>
</tr>
<tr>
<td>Domain Material</td>
<td>universal conceptual space</td>
<td>template, theoretical, schematic</td>
<td>Schematic (Schema) [E]</td>
<td>a space for domain concepts; a concept may have many possible spaces, each reflecting a certain science theory (light as wave or particle) or conception (mountain or hill).</td>
<td>- biological rank - vehicle - geopolitical unit - color - geological unit</td>
</tr>
<tr>
<td>universal concept</td>
<td>eternal, property, semantic</td>
<td>Ahistorical (Content) [K]</td>
<td>a region in an [E] space, denoting ahistorical generalizations, that are not causally dependent on a geospatial-time region; an [E] space may have multiple partitions (e.g. diverse scientific classifications).</td>
<td>- species - sport utility vehicle - country - blue - geologic formation</td>
<td></td>
</tr>
</tbody>
</table>
| Application concept | universal concept | historical, situated, pragmatic | Situational (Context) [X] | generalizations over individuals that are causally dependent on their shared history: i.e. on temporal and geographical region, place, situation, origin and interactions of individuals. | - human  
- Jeep  
- Western Country  
- Mediternanean blue  
- Dakota Sandstone |
|---|---|---|---|---|---|
| Individual [I] | single instances which maintain identity over changes of state; may include many perspectives on an [I ] within the same space-time region. | - Boyan  
- this Jeep  
- USA  
- this Med. blue  
- this rock body |
| particular individual object | individual / particular | State [S] | states of an [I] entity at a temporal point and geographical location. | - Boyan @ t1  
- this Jeep @ t1  
- USA @ t1  
- this Med. blue @ t1  
- this rock body @ t1 |

The LEKXIS framework can be viewed as a ontologic refinement of existing vertical ontology structures: i.e. from the material viewpoint it serves to partition universals into schematic, ahistorical and situational meta-concepts; from the cognitive viewpoint it partitions the notion of concepts as defined within the theory of conceptual spaces into ahistorical and situational meta-concepts, thereby giving them an ontologic interpretation, but retains the notion that relations between levels may be prototypical; and from the viewpoint of ontology kinds it partitions domain entities into schematic and ahistorical meta-concepts, and takes an ontologic approach to application ontologies by including concepts grounded in pragmatic situations rather than the narrower original notion of functions or tasks. Moreover, LEKXIS is not philosophically aligned with any of these approaches, but is instead inspired by efforts to bridge realist and cognitive directions in philosophy of mind in which a concept is a product of the relation between reality and mind, and hence can have ontologic grounding in reality while subject to (re-)identification by the mind (c.f. Millikan, 2000).

The meta-concepts (levels) are described below more fully. The descriptions are informal, and include: (1) unique defining conditions (i.e. necessary and sufficient), that enable a concept to be placed into a level (and classified according to a meta-concept), and (2) identity conditions that enable instances of a level to be differentiated.

- **Upper-level**: upper level concepts are the most general concepts as they apply to all domains. A non-exhaustive list of upper-level concepts critical to geoscience includes ‘process’, ‘event’, ‘substance’, ‘physical object’, ‘material’, and ‘quality’. Also included are their specializations and relations, e.g. a physical object might be specialized as a ‘feature’ or ‘place’, and qualities ‘inhere-in’ objects while objects ‘participate-in’ processes and events (Grenon & Smith, 2004). Upper-level geospatial concepts might be grounded in cognitive invariants such as image schemata, e.g. for ‘on’ (Frank & Raubal, 1999; Probst & Lutz, 2004). The necessary and sufficient condition for upper-level concepts is universality: they can apply to all domains. The identity condition is same schematic space, i.e. immediate instances at the next level are identical if they have the same schematic space.
**Schematic:** schematic concepts are the most general concepts within a specific domain. They derive from upper-level concepts, e.g. ‘geologic unit’ derives from ‘physical object’ and ‘earth material’ derives from ‘material’. Their distinguishing characteristic is a conceptual space (or frame), composed of dimensions for a concept (Gardenfors, 2000; Millikan, 2000). For example, the conceptual space for ‘geologic unit’ might consist of qualities for ‘thickness’ and ‘geographic extent’, their respective metrics, e.g. meters, and geospatial location as well as relations e.g. to ‘geologic process’; the conceptual space for ‘earth material’ might consist of mineral and chemical composition, texture, process, and so forth. (NADM, 2004). The necessary and sufficient condition for schematic concepts is **spatiality** and **domain-universality**: they define a space for a concept that applies across a domain. The identity condition is **same region** in a schematic space.

**Ahistorical:** common examples of ahistorical concepts are entities in classification schemes, such as types of rocks or minerals. From a structural viewpoint, they represent regions in a schematic space, e.g. ‘formation’ is a ‘geologic unit’ whose extent is mappable, ‘granite’ partitions the space of minerals, chemicals, textures, processes, etc., ‘sedimentation’ partitions ‘process’, and so on. From a philosophical viewpoint, ahistorical entities often possess unique essences (e.g. the mineral makeup of ‘granite’) and can potentially apply to any geospace-time situation (e.g. formations and granites on Mars, Earth, etc.) (Millikan, 2000). The necessary and sufficient condition for ahistorical concepts is **ahistoricality** and **regionality**: they may apply to any geospace-time situation and denote a region in a schematic space. The identity condition is **same group history**, i.e. instance identity is determined by comparing group histories (i.e. histories common to individuals within a geospace-time region).

**Situational:** situational concepts are basically historical (Millikan, 2000), which implies they can only apply to a specific situation involving a common geospace-time region and the specific interaction held in common by multiple entities within the region. In this way they are origin-dependent, as they are tied to local starting conditions and evolutionary sequence held in common by a group of individuals including: formative and affecting processes, geospatial locations, environments, and the interactions of multiple members in a temporal window. Examples include ‘species’, ‘Formation X’, ‘the granites of X’, and ‘sedimentation in a region’. Situational concepts apply to multiple entities, describing what is common or prototypical across those entities within a geospace-time region and environment. Situational qualities depend on the history of the substance they inhere-in, e.g. ‘the color of the granites of X’. The necessary and sufficient condition for situational concepts is **historical** and **multi-instantiation**: they depend on the shared history of a group of individuals that are its instances. The identity condition for instances is **same individual history** (i.e. histories unique to the individual within a geospace-time region).

**Individual:** individuals are concepts about single entities within a geospace-time region. They represent characteristics of a single entity that endure over time, with the only mandatory enduring characteristics being its identity and history. A good example of an individual is any unfragmented rock body: a specific single-body ‘Pluton C’ would be an individual, whereas the ‘suite’ it belongs to, consisting for example of a collection of plutons, would be a situational. There is a fine but clear line drawn here between individuals and situationals: the former are abstractions over a single entity in nature over
time, and the latter are abstractions over multiple natural entities over geospace-time. This implies that an individual and situational concept could apply at different times to the same rock material, for example if the material originated as a single body and was later split into several individuals. Processes, events and objects can obviously be individuals, but dependent entities such as materials and qualities can only be considered individuals in the context of the entity they depend on, for example ‘this sample of granite’ or ‘the color of this sample of granite’. Substance and process individuals differ in at least one way, in that the former generalizes states, but the latter sequences states.

The necessary and sufficient condition for individuals is singularity: they apply to single entities whose identity persists in time. The identity condition is same geospace-time position, i.e. instances are identical only if they occupy the same space-time position.

- **State:** states are not concepts *per se*, but are descriptions of the characteristics of some individual such as a physical object and its materials or qualities such as: shape, size, position, and composition. From the viewpoint of concept structure, states represent the most granular entity in a schematic space, typically a point. From a philosophic viewpoint, states represent stable descriptions of individuals over some time interval in which they are not considered to be changing, e.g. ‘Pluton C at time t1’, where the time interval is a partition of the object’s total lifetime. This assumes that even continuous processes can be separated into discrete segments of fixed description at some level of temporal granularity. The alternative is to define a type of state in which the description of an object is not stable, but where the conditions regulating description change are stable. For immediate purposes the initial assumption is sustained, recognizing that future need might require this to be modified. Then, the necessary and sufficient condition for states is object-individuality and time-indexing: they are time-indexed descriptions of individual objects.

Table 6.2 below summarizes these levels, as well as their defining and identity conditions.

<table>
<thead>
<tr>
<th>Level</th>
<th>Necessary and sufficient conditions</th>
<th>Identity conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td><strong>Universal:</strong> applies to all domains and is grounded in logical or cognitive invariants</td>
<td>same schematic space</td>
</tr>
<tr>
<td>E</td>
<td>schematic <strong>Spatial</strong> but not <strong>universal:</strong> defines a space that applies to some but not all domains</td>
<td>same spatial region</td>
</tr>
<tr>
<td>K</td>
<td>ahistorical <strong>Ahistorical</strong> and <strong>regional:</strong> applies to all geospace-time situations and denotes a region in a schematic space</td>
<td>same group history</td>
</tr>
<tr>
<td>X</td>
<td>situational <strong>Historical</strong> and <strong>many instances:</strong> applies to some geospace-time situations, and abstracts over many entities with a shared group history</td>
<td>same individual history</td>
</tr>
<tr>
<td>I</td>
<td>individual <strong>Singular:</strong> applies to single entities and is abstracted from one or more states</td>
<td>same geospace-time position</td>
</tr>
<tr>
<td>S</td>
<td>state <strong>Temporal</strong> and <strong>static:</strong> applies to the unchanging properties of an object in some timeframe and place</td>
<td>N/A</td>
</tr>
</tbody>
</table>
6.3.2 Example of LEKXIS

The example in Table 6.3 sketches an application of the LEKXIS approach to geoscience ontology. The example is fictional but representative of the domain. It involves physical objects (mountains) and related concepts (rock unit, rock materials and colors). The bold entries denote entities with schema, here called “subjects” (after Millikan, 2000), and the italicized entries denote a dependent dimension within the schema, often called a “role” in AI, or a “determinable” in philosophy when the dimension is essential to the subject (Armstrong, 1978). In representation languages, bolded subjects can be modeled as concepts and italicized dimensions as relations that connect subjects. The vertical relation between levels is assumed for now to be instance-of, and the horizontal relations at the highest level are part-of (PhysicalObject, PhysicalObject), constituted-by (PhysicalObject, Material), and inhere-in (Quality, Material) (after DOLCE; Gangemi, et al., 2003). For example, at the [X] level: the Rocky Mountains have a RockUnit part called Suite A; Suite A is constituted by a material that is a specific granite; the granite of Suite A has an inherent color that is pale-red. In more detail, informally:

- [L]: uses upper-level categories ‘physical object’, ‘material’, and ‘quality’ (inspired by DOLCE; Gangemi, et al., 2003).
- [E]: mountains have rock units that are constituted by rock matter with color.
- [K]: is a theoretical statement that mountains caused by faulting may be composed in part by a ‘Suite’ of granitic rock that is prototypically colored pink-to-brown.
- [X]: mountains within the lifespan and the geographical region of the Rockies may contain a certain ‘Suite A’ constituted by a prototypical granite with a color of pale-red.
- [I]: a specific body of rock, ‘Pluton C’, from a certain mountain, ‘Mt. Z’, is constituted by specific granites varying in [X] and having a prototypical color of light-brown;
- [S]: is analogous to [I], but for a specific point in time.

Table 6.3: example of applying LEKXIS to the ‘Mountain’ concept.

<table>
<thead>
<tr>
<th>L</th>
<th>Physical Object</th>
<th>Part-of</th>
<th>Physical Object</th>
<th>Constituted-by</th>
<th>Material</th>
<th>Inhere-in</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>Mountain</td>
<td>mt_RockUnit</td>
<td>RockUnit</td>
<td>rockUnit_Material</td>
<td>RockMaterial</td>
<td>rock_Color</td>
<td>Color</td>
</tr>
<tr>
<td>K</td>
<td>Fault_Mt</td>
<td>fault_mt_RockUnit</td>
<td>Suite</td>
<td>suite_Material</td>
<td>Granite</td>
<td>granite_Color</td>
<td>Pink-Brown</td>
</tr>
<tr>
<td>X</td>
<td>Rocky_Mt</td>
<td>rocky_mt_RockUnit</td>
<td>Suite_A</td>
<td>A_Material</td>
<td>A_Granite</td>
<td>A_granite_Color</td>
<td>Pale_Red</td>
</tr>
<tr>
<td>I</td>
<td>Mt_Z</td>
<td>mt_x_RockUnit</td>
<td>Pluton_C</td>
<td>C_Material</td>
<td>C_Granite</td>
<td>C_granite_Color</td>
<td>Light_Brown</td>
</tr>
<tr>
<td>S</td>
<td>Mt_Z_@_t1</td>
<td>mt_x_RockUnit_@_t1</td>
<td>Pluton_C_@_t1</td>
<td>C_Material_@_t1</td>
<td>C_Granite_@_t1</td>
<td>C_granite_Color_@_t1</td>
<td>Pale_Brown</td>
</tr>
</tbody>
</table>

The colors in the example in Table 6.3 are strictly [K] entities. This has to do with the relation between a quality and its value, which is independent of geospace-time. Values for qualities thus cannot and cannot be [X], [I ], or [S] entities (as shown in Gangemi, et al., 2003). ‘Light Brown’ therefore has values of “5YR”, “6”, “4” for the respective dimensions of hue, value, and saturation, as specified by the rock color chart of the Geological Society of America (Goddard, et al., 1979).
Figure 6.6: simplified example of LEKXIS in UML.
Figure 6.7: extended example of LEKXIS in UML.
Figure 6.8: OWL-FULL representation of part of the LEKXIS example.

Figure 6.6 and Figure 6.7 both illustrate and expand on the example in Table 6.3. Figure 6.6 shows a geological cross-section of a mountain, and the associated UML encoding, representing part of the example from Table 6.3 and also an added geological formation included to illustrate previous examples. Figure 6.7 also shows a partial UML encoding of Table 6.3, where additional properties have been assigned to some of the main entities, and specialization of their values and value-ranges is demonstrated. Note that although not explicitly shown, the relations and roles are instantiated from one level to the next. In addition, Figure 6.8 shows a partial OWL-FULL encoding of Figure 6.7.
6.4 Organization of LEKXIS

6.4.1 Vertical relations in LEKXIS
An open question concerns the nature of the vertical relations between levels: are they ISA, instanceof, or something more complex? This section addresses the issue and illustrates how the semantic and pragmatic viewpoints are related for vertical relations in LEKXIS.

The InstanceOf relation
From a formal viewpoint, the use of a primitive instanceof relation is often suggested between levels, though the different authors’ semantic interpretation of this relation do not always coincide:

- between [L] and [E], e.g. ‘biological rank’ is an instance of ‘substance’ (‘substance’ ← ‘biological rank’), whereas ‘kind’ and ‘stuff’ (i.e. materials) are subsumed under ‘substance’ (‘stuff’ ISA ‘substance’) (Millikan, 2000);
- between [E] and [K], e.g. ‘biological rank’ ← ‘species’ (Sowa, 2000, p. 32);
- between [K], [X] and [I ], e.g. ‘species’ ← ‘human’ ← ‘Chris’ (Guarino & Welty, 2002);
- between [E] and [S], [K] and [S], as well as [X] and [S], e.g. geospatial objects, at a time instant, and their corresponding universal entity, based on an a priori universal/particular distinction (Bittner, et al., 2004);
- between [X] and [I ], e.g. ecological unit classes and ecological units (Sorokine, et al., 2004);
- between [I ] and [S], e.g. persistent individuals and their state at some point in time (Heller, et al., 2004).

In all these levels, the instanceof relation honors only some of the effects of concepts and disregards others such as instance revision, i.e. instantiation; it also ignores their origins. It is therefore only partially pragmatic. A fuller pragmatic account of vertical relations would also include the clusterMembership relation, which relates a cluster and hence a concept with originating individuals. Both relations are shown in Figure 6.9: the instanceof relation uses a dashed open arrow, and the clusterMembership relation uses a dashed closed arrow. Note that the cluster and class overlap for a situated concept but are not identical, as the concept’s origins include contrasts as well as instances.

Figure 6.9: parts of a cluster and class for a situated concept; open arrows denote instanceof relations, and closed arrows denote clusterMembership relations.
The ISA relation

ISA is commonly interpreted intensionally as the relation between a generalized concept and a specialized concept. In LEKXIS this interpretation can be maintained by structuring ISA as the relation between a space and subspace within each level, such that subspace ISA space (after Ahlqvist, 2004). ISA therefore applies within levels, unlike instanceOf which applies between levels:

- within [E]: between schema and sub-schema, such that schema ISA sub-schema, reflecting the convention that reducing dimensions of a schematic concept represents a generalization, e.g. ‘Munsell Color’ ISA ‘Color’ and ‘RGB Color’ ISA ‘Color’;
- within [K]: between regions and sub-regions in thematic space, such that sub-region ISA region, e.g. ‘navy blue’ ISA ‘blue’, ‘monozogranite’ ISA ‘granite’;
- within [X]: between histories and sub-histories of groups of individuals, e.g. ‘some sub-species’ ISA ‘some species’. When causal links are not pertinent, then regions and sub-regions of geospace-time can be used, e.g. ‘WWII Jeep’ ISA ‘Jeep’, ‘granites of the Canadian Rockies’ ISA ‘granites of the Rockies’;
- within [I]: between histories and sub-histories of individuals, e.g. ‘the quiet Boyan’ ISA ‘Boyan’, or ‘this rock body of the Dakota Sandstone at some time interval’ ISA ‘this rock body of the Dakota Sandstone’. Each of these concepts represents a generalization over multiple states of an individual;
- and [S]: states are analogous to points in space and are therefore without subspaces.

The subspace relation can also be used to characterize the semantic relation between ontologic, perspective and version concepts. Recall that ontologic concepts are defined to be concepts about nature as revealed by science, and that perspectives and versions are defined to be alternative scientific views. Perspectives often carve the many inherent qualities of an ontologic concept into disjoint partitions, according to scientific methodology, and versions are then further refined here as specializations of perspectives in which epistemic qualities are added to perspective concepts. For example, geologists can describe a geological unit from a fieldwork perspective using a certain group of qualities that are disjoint from those provided by a satellite for the same unit, and different geologists can have divergent fieldwork versions for the unit as a result of varied purpose, training, and so forth. Another important distinction is the contrast between ontologic, perspective and version concepts and another type of concept called a role. Roles are contingent on uses carried out by humans and others, e.g. a geological unit or ecological unit, and can be contrasted with the contingent use of the region as a ‘nuclear waste disposal site’.

From these definitions it can be concluded that the semantic relation between ontologic, perspective and version concepts at the same abstraction level is ISA. The following reasons support this conclusion: firstly, following Welty and Guarino (2001), no ISA constraints are contravened, e.g. no perspective concept is a role, because the qualities are not contingent but are necessary in all possible worlds; secondly, if at each level the perspective concept is considered a union of all possible version spaces, and the ontologic concept a union of all possible perspective spaces, then the ISA relation reverts to the original subspace relation in which versions are subspaces of perspectives, which are subspaces of ontologic concepts. Examples of perspective-ontologic concept relations among qualities include:
• within [K]: ‘Munsell blue’ ISA ‘blue’;
• within [X]: ‘pink of Formation X from fieldwork’ ISA ‘pink of Formation X’;
• within [I]: ‘pink of rock body 1 from fieldwork’ ISA ‘pink of rock body 1’.

These examples for [X] and [I] assume that the colors of rocks could in addition be obtained by methods/perspectives other than human fieldwork, such as remote photography as deployed, for example, by the Mars Pathfinder mission.

Now, can the ISA relation be applied between the ontologic concept of a state, its various perspectives, and its versions, for example between ‘Rock Body X1@t1’, ‘Rock Body X1@t1 from fieldwork’, and ‘Rock Body X1@t1 from fieldwork of geologist 1’? If the ontologic state is indeed an abstraction from versions within perspectives, as indicated by the pragmatic viewpoint taken here, then it would follow that relations between these concepts are also ISA.

Finally, like the instanceOf relation, the ISA relation is strictly semantic: it is only concerned with the products of abstraction and not the pragmatic process by which abstraction occurs. Given that ISA represents concept abstraction within a level, i.e. between concepts, then the associated pragmatic relation is blendMembership. For example, in Figure 6.10 below, a fragment of an earlier diagram (Figure 5.8) is extended to include both the semantic ISA relation and the pragmatic blendMembership in the abstraction of a situated concept from one fieldwork perspective and two geologist’s versions.

![Figure 6.10: semantic ISA relation (closed arrow), and pragmatic blendMembership (open arrow) relations between situated concepts shown as a UML diagram.](image)

### 6.4.2 Horizontal Collections: Extension, Model, Possible Model

A collection of entities within a level may be considered: (1) a class: a collection of instances of one higher level concept; (2) a model: a collection of instances of several concepts and relations from a higher level (e.g. the collection of concepts at any level in Table 6.3 represents a geologic model for that level of abstraction).

Common types of natural science models include:

- collections of [E] entities are often the set of predicates in general domain theories; e.g. a theory of certain mineral formation;
- collections of [K] entities represent classification schemes; e.g. types of mineral deposit models;
collections of \([X]\) and \([I]\) entities can represent a conceptual model for an area, involving places, spatial relations, time relations and places but not geometries; e.g. a conceptual model for a particular mineral deposit;

collections of \([S]\) entities are physical models, involving absolute space and time locations; e.g. a 3D model of a mineral deposit consisting of a collection of geometries.

6.5 Application of LEKXIS

6.5.1 K-R Languages

Encoding of concepts at the various abstraction levels can be accomplished by languages with expressivity that allow entities to be both classes and instances, e.g. as in the rudimentary UML and OWL encodings above. Some authors have investigated solutions that preserve first-order encoding in such languages, incurring some constraints on reasoning (e.g. Noy, 2004; Welty, 1998). These and other aspects involved in the engineering of LEKXIS require much more attention.

6.5.2 Problems that can be addressed by LEKXIS

The problem most readily addressed by LEKXIS is confusion in semantic granularity. This problem most commonly occurs when a request is made to data sources for information at a certain implied semantic granularity, and results are returned at different granularities. For example, when individuals or states are desired but the system returns situationals, i.e. situated concepts \([X]\), or ahistorical concepts \([K]\), i.e. returning the definition of ‘granite’ from a classification scheme instead of the description of the granites for a specific polygon on a map. This confusion often happens unknowingly because the levels are not distinguished by the system. It leads to the use of a description that is too general, or one that is too specific, for some scientific task.

Semantic granularity in LEKXIS can be viewed from the perspective of a semantic cube possessing three dimensions, as shown in Figure 6.11: geospatial regions, time regions, and individuals. The time axis is measured in units of time points: “one time” refers to a single point in time, “many times” refers to a continuous time interval containing several ordered time points, and “all times” refers to all ordered points in time until the present. The geospatial axis is measured in units of spatial regions where a region has a single surface boundary that is unbroken and can be registered to some geospatial coordinate system. “One place” refers to a single region in the coordinate system, “many places” refers to multiple distinct regions in the coordinate system, and “all places” refers to all possible regions in the coordinate system. The individual axis is measured in units of individuals, where “all individuals” refers to the collection of all individuals present in some geospace-time region.

Volumes in the semantic cube then designate the range in which meta-concepts can apply. For example, states occur for a single individual, at one place and time. Individuals are single entities, but occur at multiple times at one place at each time. Situationals are realized across multiple individuals, times and places; and ahistorical, schematics and uppers can apply to all times and places, to many but not all individuals. An exception to this scope for uppers is a concept such as ‘entity’ that applies everywhere, every time to everything; this concept will be ignored for present purposes. Then, ‘granite’ \([K]\) can apply to some body of rock material at any time and place; ‘the granites of Suite A’ \([X]\) applies to multiple individuals at multiple
places over some time interval; ‘the granites of Pluton C’ [I] apply to one individual located in possibly multiple places over a time interval, and ‘the granites of Pluton C @ t1’ [S] apply to one individual, at one time and place. Semantic granularity thus increases in the semantic cube from bottom back right to top near left. Note the semantic cube provides some but not necessarily all of the defining conditions described above for a meta-concept, i.e. defining conditions for situated concepts additionally consist of a shared history amongst instances.

[Figure 6.11: semantic granularity in the semantic cube.]

The semantic cube can now be used to define semantic confusion in terms of semantic granularity. Semantic confusion refers to the inappropriate substitution of a concept at one level of granularity with a concept from another level, for some task. Such confusion can be illustrated in a confusion matrix, in which the impact of substituting amongst entities with different granularities is made apparent. The confusion matrix shown in Table 6.4 can be read from left to right in each row: the meta-concept to the left of the row represents the desired granularity for some task, and each cell in the remainder of the row, to the right, indicates the nature of the confusion resulting from substitution of another type of meta-concept (indicated by the column). Confusion is indicated as being too specific, i.e. “over-granular”, or too general, i.e. “under-granular”, as determined by the change in granularity in the semantic cube when comparing substituted to desired meta-concept: i.e. over-granularity occurs when the substituted meta-concept is more granular than the desired meta-concept,
and under-granularity occurs when the substituted meta-concept is less granular than the desired concept.

Table 6.4: semantic confusion matrix.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>State [S]</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Individual [I]</strong></td>
<td>under-granular: in time and place</td>
<td>under-granular: in time, place, individuals</td>
<td>under-granular: in time, place, individuals</td>
<td></td>
</tr>
<tr>
<td><strong>Situational [X]</strong></td>
<td>over-granular: in time</td>
<td>over-granular: in place, individuals</td>
<td>under-granular: in time, place, individuals</td>
<td></td>
</tr>
<tr>
<td><strong>Ahistorical [K]</strong></td>
<td>over-granular: in time, place and individuals</td>
<td>over-granular: in time, place, and individuals</td>
<td>over-granular: in time, place, and individuals</td>
<td></td>
</tr>
</tbody>
</table>

Example of semantic confusion between states [S] and individuals [I]

An example of over-granularity in the time dimension occurs when a state is substituted for an individual (third row from top, second column from left), because the state refers to one time point but the individual ranges over many time points—this would reflect a situation where it is important to know the properties of an individual that apply during its full lifespan, rather than during some specific interval of its life, e.g. knowing for land-use planning purposes the mineral potential, forest fire potential, and so forth, for some geographical region rather than the actual acreage being mined or under fire in that region at some time. An example of under-granularity is the reverse situation, when a state is desired but an individual is used, such as in an emergency response scenario in which potential values for mineral and forest fire occurrence are used for the region, instead of the actual values of mining and burning at the time. Apart from confusion in time, the confusion matrix also shows over-granularity and under-granularity in terms of number of places and individuals, in that the substituted concept can apply to too few/many places and/or individuals.

Example of semantic confusion between individuals [I] and situationals [X]

A good example of where semantic confusion would be costly and dangerous is hazard risk assessment. In hazard risk assessment insurance companies are beginning to use underlying geologic conditions to determine policy strategies, and an important aspect of these conditions is rock material. For example, the intensity of a seismic wave at the ground’s surface depends on the propagation characteristics of the rock material. Likewise, landslide potential involves the degree of consolidation of the material, among other factors such as slope of the landscape. A major source of rock material descriptions is the geologic map,
where they are found in two places: somewhat rarely in descriptions of individuals (acting as proxies for the current state), typically shown as polygons on the map, and more commonly in descriptions of situationals, typically found in the map legend, and which are prototypical for the many individuals that are instances on the map. Importantly, due to physical heterogeneity, the situational prototype and the individuals may vary significantly, affecting calculations of the likelihood of a hazard occurrence, its intensity and thus also estimates of potential damage, if one is substituted for the other.

These differences can be enormous in terms of costs and losses. Yet, the scenario is realistic because geologic map databases do not commonly distinguish between situationals and individuals—they usually store individuals with assigned rock material descriptions, which are sometimes specific to the individual, but usually inherited from the prototype. The point is that error estimates on any risk calculation would benefit by knowing if the rock material descriptions are local or prototypical, as well as the thematic distribution of instances of the prototype, in order to gauge reliability and uncertainty of the calculation. Unfortunately this rarely occurs because the types of entities are not formally discriminated in geological systems, nor is their distribution specified.

**Example of semantic confusion between individuals [I] or situationals [X], and ahistoricals [K]**

The situation is compounded when the prototypical or individual rock material descriptions are sparse, and simply refer to an ahistorical concept from a classification scheme, e.g. where an individual polygon [I] or situational map unit concept [X] is simply described as constituted by sandstone, rather than a sandstone with such and such properties. Then the rock material properties must be inherited from the ahistorical [K] description, which do not describe the particular material as accurately as descriptions at situational [X] or individual [I] granularities, resulting in potential loss of accuracy in any subsequent calculation.

Rock material properties are an important ingredient in many other societal applications that involve the landscape, such as engineering (e.g. constructing infrastructure), groundwater (water supply), vegetation (agriculture), environment (animal habitats), and energy (petroleum). Disambiguating abstraction levels of rock materials then has significant implications for societal decision-making, which are only beginning to be recognized and whose effectiveness can be improved by such disambiguation.

In summary, the levels of abstraction exhibit semantic granularity that is linked to geospace-time granularity. Values of properties at each level can be seen as generalized from levels below, such that appropriate use of a concept for some critical task requires knowledge of its geospace-time extents and their fitness to a use-case. Usage of a concept with inappropriate granularity can lead to disastrous and costly consequences.

**6.6 Conclusions**

This chapter identifies six levels in the vertical structure of ontologies for substances, and provides examples from geoscience. The levels/meta-concepts consist of: [L] formal concepts, [E] domain schemas, [K] ahistorical entities that are regions in a schema space, valid across all time and geospace, [X] situational entities that are regions in a schema space, historically-causally tied to a situation in a geospace-time region, [I] individuals that are single entities with persistent identity, and [S] states, that are time-indexed descriptions of
individuals. Examples from the earth sciences demonstrate how these levels may be applied, including prototyping in the UML and OWL representation languages.

One consequence of these levels is improved definition of some notions introduced in earlier chapters: situational knowledge can be now redefined as consisting of ontologies with contents from the situational, individual or state levels, and theoretical knowledge can be now redefined as consisting of ontologies with contents from the remaining more abstract levels, i.e. the ahistorical, schematic and upper levels.

Another important benefit of the LEKXIS framework is increased semantic granularity. This implies improved differentiation between types of concepts, impacting how those concepts should be used in various scientific and societal decision-making tasks where it might be inappropriate to substitute concepts from one level with concepts from another. A second benefit is increased semantic precision, in which meta-constraints at various levels have been explicated and refined. A third benefit is implied but not demonstrated. It involves improved ontology design, in which the meta-concepts and meta-constraints provide guidelines to ontology structure. This benefit has a corollary: improved design based on principles resonant with the geoscience domain should lead to improved knowledge capture. Potential future work involves development of the meta-constraints into rules that can guide the hierarchical placement of concepts in ontologies, supplementing other meta-properties that consider the hierarchical ISA relation primarily (e.g. Guarino & Welty, 2002). Once these are in place, it is conceivable to use them to help advance knowledge capture. Future work must also include implementation of the LEKXIS framework in practical geoscience applications, probably within some cyberinfrastructure.

Finally, a somewhat implied consequence of the LEKXIS framework is the raised stature of the situated concept [X] in ontologies, and thus the rise to prominence of the role of the geographic in ontology structuring; this is in addition to its already recognized role in contributing geospatial and geographical concepts to the content of ontologies. In explicating and framing situated concepts, which are essentially geo-pragmatic, in terms of their role in ontology structure, this chapter addresses the first geo-pragmatic challenge and the fifth objective, both of which aim to explore the impact of geo-pragmatics on semantics.
CHAPTER 7: CONCLUSIONS

7.1 Summary and conclusions

7.1.1 Assumptions
This dissertation examines the representation of geoscientific ontologies from the viewpoint of how geoscientific concepts are developed, evolved, and contextualized in the process of doing science—i.e. from a geo-pragmatic viewpoint. Some important assumptions related to the representation of ontologies immediately arise from this viewpoint, and guide this work throughout: the ontologies of interest are those that contain geoscientific concepts, where such concepts are abstractions over geoscientific entities such as natural objects, process, properties, etc.; the concepts and entities are products of scientific inquiry and have claims to some existence in natural reality, albeit through the lens of certain modes of scientific detection that assumes a hybrid semantics linking reality and mind; and because scientific knowledge often evolves, fragments and conflicts, it is typically contextualized, most notably along dimensions for origin, effect and use, in which humans, machines and nature are variably producers and consumers of artefacts subject to such contextualization. These boundaries are introduced in chapter 1, developed in chapter 2, and are applied in the remainder of the dissertation.

7.1.2 Objectives and Challenges
Five objectives and three related challenges are addressed by this dissertation, and are reviewed in this section. The first objective aims to identify semantic and geo-pragmatic challenges to geo-interoperability (chapter 2), and the remaining objectives set out to explore the identified challenges. The second objective aims to develop a conceptualization and formalism for representing geo-pragmatics (ch 3); the third objective then tests the fit of the conceptualization to some empirical data, to address the empirical challenge (chapter 4); the fourth objective tests the formalism using the empirical data, to address the schematic challenge (chapter 5); and the fifth objective explores implications of the conceptualization to ontology design, to address the semantic challenge (chapter 6).

The first objective, identifying challenges to representing geoscience knowledge, is met in chapter 2 through a review of approaches to information interoperability, ontologies and geoscientific knowledge, in which the nature of geoscientific knowledge is shown to constrain the structure of geoscientific ontologies. In particular, geoscience knowledge is characterized as being fragmented by methodological and thematic perspectives that evolve and compete to create versions of both theories and geographical models. Representational constructs for such knowledge must then manage contexts to account for the lineage of theories and models, to enable testing and replication of results, and to account for competing scientific ideas and their convergence into (temporary) unified bodies of knowledge. In response, geo-pragmatics is proposed as a key theoretical mechanism for representing contextual elements of concepts. It consists of a two dimensional matrix-like structure consisting of actual origin, effect and use aspects of the concept on one axis, and the human, machine and natural domains on the other axis. The key idea is that a concept may be supported by interactions within domains and between domains, such that e.g. human interpretation can derive from machine representations of nature (e.g. remote sensed images), from direct observation of nature, or from other humans. However, ontology representations
do not inherently support representation of such context, raising three challenges: (1) the semantic contents of ontologies are challenged by the proposal of a novel type of pragmatic-driven concept, the situated concept, which is unconventionally dependent on the historical interaction of its instances; (2) the schematic structure of ontologies are challenged particularly by the need to include functions inherent in scientific inquiry in addition to the products of such inquiry; and (3) these ideas need to be tested empirically to evaluate the fit of theory to scientific practice.

The second objective, developing a theory for geo-pragmatics, is met in chapter 3 with the development of a conceptualization for geo-pragmatics that is described using a logic-based formalism. The conceptualization is based on the matrix-like structure for geo-pragmatics developed in response to the first objective, supplemented by a more granular categorization of primitive geo-pragmatic entities including concepts, individuals, states, objects and properties. The conceptualization also adapts a model for scientific reasoning, leading to five primitive pragmatic functions that model the results of concept discovery and evolution: clustering, blending, individuation, classification, and instantiation. The structure that describes the interaction of the primitive functions and entities is called a pragmatic square. It is particularly relevant to geoscience because it provides a model for how geographical objects and concepts are inferred from properties, a common occurrence in geographical discovery in which the object and its classification are not known beforehand but are discovered in tandem. A logic-based formalism is outlined for the conceptualization and a representative example of geologic region discovery is demonstrated using logic.

The third objective sets out to empirically test the conceptualization of geo-pragmatics in chapter 4, and by doing so also meets the third geo-pragmatic challenge of verifying that the conceptualization has a good fit to existing data and geoscientific practice. The field data collected by three field geologists collaborating to develop a geologic map for a geographic area are analyzed. In this analysis clustering trends in the data are compared within geologists over time, and between geologists over time, to reveal that the situated concepts developed by the geologists during the mapping exercise are influenced by natural, human, theoretical, situational factors, and characterized by varying uncertainty and change. This further implies the meaning of such concepts is related to these pragmatic elements, substantiating the need to introduce them into concept representations. Of particular significance are the implications for map development and map quality assessment, in that the techniques developed here enable map concepts to be assessed for certainty, coherence and completeness, to help determine the trustworthiness of a map and mapper.

The intent of the fourth objective and second challenge, both met in chapter 5, is to test the formalism against real data, by developing a schema from the formalism and populating it with data from the empirical study. The choice of building a database schema against geoscientific data raises an interesting issue in database modeling, and conceptual modeling in general: the design of such schema is predicated on pre-determined and static concepts, but situated concepts such as those developed in geologic mapping are neither. The solution adopted here models situated concepts directly as entity in the schema, which then requires the remainder of the geo-pragmatic machinery to be added for a complete representation. This design was tested by building a prototype geo-pragmatic database schema in a commercial GIS, populating it with data from the empirical study, and visualizing the results in geographic space. Of particular interest are visualizations of the observations deemed by
the empirical study to have been critical to concept discovery. Viability of the schema is demonstrated by its successful implementation in the prototype, and by several significant third parties who experimented or deployed various parts of the schema. Most of these implementations deployed a conceptual network version of the schema, i.e. including concepts and related instances, but ignoring the pragmatic aspects. The benefits to these third parties included simpler and more robust database design, and improved fit of schema to data and to the scientific process, offset by increased query complexity. The schema is particularly suited to use in situations where scientific knowledge is being discovered or evolved, and where tracking discovery and evolution is important.

The intent of the fifth objective, and first challenge, both met in chapter 6, is to explore the impact of situated concepts on ontology structure. Although such concepts have been loosely recognized in ontology design for some time, they have not been clearly identified and their role in ontology structure has not been well explicated. A framework, called LEKXIS, is therefore developed which proposes six levels of abstraction for concepts, including one for situated concepts, and provides informal grounds for distinguishing concepts at each level. Intriguingly, it is the fundamental geographical nature of situated concepts that distinguishes them—their dependence on geospace-time relations amongst instances. Viability of the framework to geoscience ontology is shown via example implementations that use the UML and OWL ontology languages. The framework faces known and considerable engineering challenges, as well as further testing in serious geoscience applications to gauge its suitability.

7.1.3 General Conclusions

The representation of geo-pragmatics is shown to be realistic, viable, and valuable. It is realistic in the sense that the conceptualization fits the geographical knowledge discovery exercise of a group of geoscientists, one that is fairly representative of many mapping activities in which regions are discovered. It is viable because the conceptualization can be analysed using novel methods of concept analysis, and because the formalism is shown to translate to a schematic structure that can be implemented in a variety of computing environments, albeit with some engineering trade-offs (see section 7.2.4 below). Its value is apparent in its impact on several third parties, and in the insights gained about geographical knowledge discovery in general, geoscientific mapping and mappers in particular, and in the geographic contribution to ontology structure.

7.2 Contributions, limitations and future directions

7.2.1 Semantics and geo-pragmatic representation

Taking a pragmatic viewpoint on ontologies is not unique, as cognitive and social approaches are often dynamic or emphasize context. Nor is it unique that the focus here is on geospatial entities, for these also are well represented amongst those approaches applying human-oriented perspectives to geospatial representation in machines. Significantly, such approaches are particularly pragmatic because they are grounded in primitive functions and biological hardware that give rise to socio-cognitive knowledge. But, they are rarely, if at all, grounded in any computer-friendly theory that models knowledge discovery, evolution, or context, and when they are headed in this direction, the theoretical foundation is strictly cognitive or social, and not scientific. Opposed to this are more traditional approaches in
which ontologies are engineering artefacts constrained by reasoning formalisms typically implemented in logic, and/or in which ontologies are representations of reality, and scientific product, but ignorant of scientific process.

This dissertation responds to the obvious gaps. At the core of this work is the idea that the primitive functions and associated entities identified herein represent a science-based meta-structure for concept representation in ontologies. The theoretical significance of the structure and associated conceptualization is in its integration of the representational aspects of the cognitive and realist traditions, from a scientific viewpoint. This integration involves, on the one hand, inclusion of the effects of concepts which denote the products of science of concern to realists, as well as on the other hand, the origins of concepts which involves the epistemic and other factors related to discovery of concern to socio-cognitivists.

Another related representational bridge between the two camps is the nature of a concept: in some cognitive and social views concepts are not primitives. In these views the primitives are essentially amorphous mental entities or social byproducts that take on various roles, including that of concept, as required in situations (e.g. Barsalou, 1999)—they are essentially nodes, or *signs*, in an association network that are adapted to circumstance. In stauncher realist circles, the nearest notion to an associative node is that of a universal, which is a repeatable pattern found in nature but which is not mutable, and which obeys ontological principles that carves entities into distinct categories, e.g. objects, properties, materials, processes, etc., that can never switch roles because of their ontological primacy. The fact that roles are often switched in day-to-day conceptual modeling is then under this view simply imprecise practice. Although these views seem completely opposed, they can however be bridged for ontology representation purposes with some comprise. Thus, the middle position achieved here has elements of both, in its structure and contents, and also in terms of its underlying principles which adopt a hybrid position that allows for qualified access of reality from mind, and thus for the scientific knowledge of reality.

Structurally important ontologic distinctions such as object, property, concept and individual are preserved in the middle position described in this dissertation. The impetus for this involves the recognition by both realist and socio-cognitive camps of these types of entities. Although the chicken-egg debate surrounding them is likely irreconcilable (as it has been for millennia), i.e. viewing ontologic primitives as the result of socio-cognitive activity versus viewing socio-cognitive activity as exemplifying *a priori* ontologic primitives, it is sufficient here as a start, for representation purposes, to accept the structures held in common regardless of their interpretation and whether instances of these constructs in nature are ‘TRUE’.

Then, the socio-cognitive notion of association networks is added to link these common distinctions, because it reflects scientific practice, but this is accomplished via an ontologic twist in which the links are categorized as origins, effects and uses. Hence the links of a causal association network for concepts are given a pragmatic interpretation grounded in some foundational functional notions.

This leads to a distinction between two types of association networks, pragmatic networks whose links denote functional interactions of concepts characterized as origins, effects, and uses, and a subset of this called a conceptual network which is interested only in certain effects and semantic relations, where the links are characterized by semantic relations such as
ISA and instanceof. In computer science, many conceptual modeling environments (e.g. E-R, UML), and representation languages (e.g. OWL, KIF) are in fact either conceptual networks with a narrow range of links, or associative networks in which the links can be specialized to represent a wide variety of relations and semantic interpretations, including those that are pragmatic (e.g. RDF, semantic networks, neural networks, conceptual graphs (Sowa, 2000)). Thus, pragmatic networks as described herein lie between these two, in that they consist of associative networks in which nodes are concepts containing pragmatic as well as semantic links (i.e. origin, effect, use) and where conceptual networks are the subset that contain semantic links mainly to model certain effects of concepts. In the end it is through such pragmatic networks that geo-pragmatics is represented. But, there exists no clear diagrammatic notation or convention for illustrating such networks (in spite of some ad-hoc attempts made in this dissertation), a clear detriment to their use. Presenting and using pragmatic networks is a future challenge.

7.2.2 Methods for evaluating map quality

An important contribution of this work to GIScience is the development of novel quantitative measures for map quality, including the accuracy, coherence, completeness, etc., of concepts used to classify map features. This contrasts with measures for the quality of the assignment of concepts to features, or of the features themselves (e.g. Aspinall & Pearson, 1994; Congalton, 1991; Goodchild, et al., 1992; Schaffer, 1993). The first of these is often called classification accuracy and considers whether a feature is properly classified. The second of these is often associated with “feature-level” metadata, and is intended to qualify the feature’s properties, such as uncertainty in the position of its boundary. From the geo-pragmatic perspective both of these represent efforts toward a qualitative assessment of two of five pragmatic functions, classification and individuation, respectively. One can then imagine issues of quality for the remaining functions of clustering and blending (or concept development), as well as instantiation (or property development), though these are rarely, if ever, considered in quality assessments, much less quantitatively or visually measured.

Therefore, in this dissertation, several novel geocomputational and visual metrics and methods for evaluating concept development are developed and tested. These involve metrics and methods for the comparison of geologists’ data clusters over time for a concept, and reveal that geologists’ views of a concept vary during and at the end of the mapping campaign. The potential for re-using these metrics and methods for overall map assessment is significant, as are the implications of the assessments which suggest they could inform collaborators about conceptual trends, as well help choose experts for certain tasks. For example, measuring the consistency of the various rock-related classifications made by the Mars’ robots would help guide the data collection through insights into physical heterogeneity; it would also help detect faulty equipment. Indeed, the metrics and methods are applicable in any situation where time-indexed conceptual knowledge is being abstracted from source data; they are not limited to the geographical domain.

Further work is required to bring mathematical rigor to the MMD statistical techniques employed, as the formal shortcomings and benefits of the approach are not developed nor well recognized. Any serious use of the MMD method would require a thorough analysis of its boundaries, and axiomization of its assumptions and properties. Finally, there is considerable opportunity to develop tools that deploy the techniques built around MMD, and in particular to enable dynamic visualization over time, perhaps using the schema from
chapter 5 as a data structure. Indeed, these are probably a necessity if the benefits listed above are to be realized (Gahegan & Brodaric, 2002).

### 7.2.3 Empirical studies of geographical knowledge discovery

A surprising result of the work carried out here is the insight gained into geographical knowledge discovery. As mentioned before in chapter 4, most studies of geographical concepts occur in controlled laboratory environments and are aimed at uncovering the character of existing concepts for landscape features, not regions, or they strive to uncover the essentials by which geographical concepts are placed into classes. In contrast, this work studies how geographical concepts are developed to begin with, and reveals the pragmatic nature of the mapping process of a team of geologists. Possible uses for the knowledge are given above.

Potential future work involves tracking the development of prototypes. The situated concepts, described in the map legend, represent prototypes of the observed data as agreed to by the mappers at the conclusion of the mapping exercise. Knowing this final prototype enables inferences about the nature of the concept to be made based on the distance of the cluster to the prototype. A wrinkle in this current study is the fact that the final prototype is assumed as being given from the outset, when in fact it is also developed during the course of the mapping. An interesting follow-up study could then study prototype development, in comparison and contrast to cluster development. This could be realized quite simply by having geologists indicate at each site a degree of typicality for the site, i.e. how similar is the site to the prevailing prototype? A transition in data for typical sites would then indicate a shift in the prototype; correlation between the evolving prototype and the clusters or final prototype could then also be measured and visualized.

Potential future work also could include a study of the influence of frequency of observation on clustering and concept development. In this work duplicate data were not removed, but were included in the study. They had the effect of focusing clustering in the thematic areas occupied by the duplicates. The argument for including duplicates suggests that the sampling density is relatively sparse making all samples potentially significant, and that each duplicate acts as confirmation for the inference by which the region is discovered. But this argument is an untested hypothesis that begs to be explored.

Another weakness of the current study is the difficulty in getting (dis)confirmation for some of the results and interpretations. This is mainly due to a well-recognized issue in studies of scientific discovery in which *in situ* studies are preferred to the mining of scientists’ memories (Klahr & Simon, 1999) Thus, in future work mappers could be shadowed *in situ* and questioned about their conceptualizations, and the results compared with the quantitative analysis, e.g. to help pinpoint when instantiation and individuation are occurring. *In situ* studies could further be supplemented with *post hoc* exit interviews. This mixture of improved prototype tracking, *in situ* and *post hoc* corroboration should increase the validity of the results and interpretations, and potentially lead to more fascinating outcomes.

### 7.2.4 Geographical ontologies and GIScience

This dissertation makes two theoretical contributions to the study of geographical ontologies. The first contribution is the development of a conceptualization and computable formalism
for geographic concept discovery. The second contribution is the identification of situated concepts and the resultant ontology framework called LEKXIS.

The first contribution, a geo-pragmatic conceptualization and computable formalism for geographic concept discovery, is significant because geographical knowledge development is rarely studied from a foundational perspective. Cognitive studies tend to concentrate on characterizing the human classification (or ‘categorization’) function, i.e. how humans place an entity into a class (or ‘category’), while ontologic efforts ignore the function and focus on formalizing the underlying entities and relations. In contrast, geo-pragmatics strives to capture the full cycle of knowledge discovery and evolution, by building on ontologic primitives such as property, object, cluster, class, etc., and formally connecting these to reflect the fact that regional-scale geographic objects and concepts are often developed from observations of properties.

The main impact of the conceptualization and formalism then rests in their ability to aid discovery of new knowledge, through operationalization of the pragmatic functions, and in their use as an archival platform for old knowledge and its pragmatic context. They could therefore play a key role in shaping scientific cyberinfrastructures by providing some theoretical footholds upon which to architect workflows and build tools that help discover new knowledge. Apart from impacting cyberinfrastructure design and activity, they could also impact the content of ontologies by providing a conceptual and formal template for recording and tracking contexts, including the context of discovery (i.e. origins).

Both of these areas of potential impact represent possible future applications for geo-pragmatics. For these impacts to be realized some theoretical holes need to be at least considered, if not plugged. One of these holes considers the grounds for applying the pragmatic square. In chapter 3, it was decided for convenience to restrict application of the square to generalizations between abstraction levels, and treat all generalizations within levels as a special case of blending. For example, the current state of some natural object as agreed upon by a community of scientists (i.e. its ontologic state) may be a generalization of several competing versions promoted by sub-communities, and each competing version may itself be a generalization of views held at different time points. Because each view of the state of the natural object necessarily possesses properties, the pragmatic square could apply, as shown in Figure 7.1.

![Figure 7.1: applying the pragmatic square within an abstraction level.](image)
This raises questions about the meaning of the generalization mappings \( F, G, \) and \( f, g, \) originally described in section 3.2 and Table 3.2. These mappings are associated with the clustering functions for properties \( (F) \) and objects \( (G) \), and with the functions for blending \( (f) \) and individuation \( (g) \), by which entities are derived from sources. Now, are \( F, G \) semantically different when applied within levels rather than between levels, and if so, how? Are mappings \( f \) and \( g \) dependent on the role being played by the concept, as abstraction and instance, respectively? These questions remain unanswered, but are of theoretical concern.

Further work is also required on the formalism for the pragmatic square and some consequent implications for modeling geographical knowledge discovery. Along these lines, Werner Kuhn (2002) asks the question whether a geographical category is an algebraic category? If so, then the apparatus of algebraic formalism can be applied to cognitive (geographical) concepts. The impetus for doing so would be theoretical elegance, and increased computational traction. Kuhn’s question can be rephrased here for similar purposes: “is a pragmatic concept an algebraic category?” Then the pragmatic square can be immediately seen as being somewhat categorical, in that it defines certain primitive entities and mappings for concepts. But does it obey the axioms of algebraic category theory? The answer to this question involves applying the category axioms to the pragmatic square, and testing the results for validity against independent domain theory. The challenge rests with the second of these tasks, in light of the absence of a comprehensive theory for geographic knowledge development. The issue is exemplified in Figure 7.2, in which the pragmatic square is tested against the composition axiom, using LEKXIS as the domain theory.

The composition axiom asserts that mappings are transitive, which implies the pragmatic square can be stacked vertically such that, given concepts for individuals \( I \), situationals \( X \) and ahistoricals \( K \) from LEKXIS, and property clustering functions \( F_1, F_2 \), (both instances of \( F \)) as well as object clustering functions \( G_1, G_2 \) (both instances of \( G \)) from the pragmatic square, then:

\[
\begin{align*}
\text{Eq. 7.1:} & \quad F_1: I \to X \quad \text{and} \quad F_2: X \to K \quad \Rightarrow \quad F_1 \cdot F_2: I \to K \\
\text{Eq. 7.2:} & \quad G_1: I \to X \quad \text{and} \quad G_2: X \to K \quad \Rightarrow \quad G_1 \cdot G_2: I \to K
\end{align*}
\]
In plain language, Eq. 7.1 and Eq. 7.2 state that if ahistorical concepts \( K \) are abstracted from situated concepts \( X \), which are abstracted from individuals \( I \), then ahistoricals \( K \) must be abstractable from individuals \( I \). This makes intuitive sense: the ‘species’ concept must be abstractable from ‘human’, ‘cheetah’, etc., as well as from ‘Chris’, ‘Tigger’, etc. A complete categorical account of pragmatics would thus need to apply the categorical axioms to each of the five pragmatics functions and test their validity against a domain theory, which itself would likely require development or modification.

The second contribution of geo-pragmatics to geographical ontology is the identification of the intrinsically geographical situated concept. The situated concept serves to formally identify a type of concept for those entities commonly described on map legends and used to classify the geographical regions that are frequently portrayed on maps. Situated concepts therefore could help bridge digital cartography with ontologies, and pave the way for several areas of potential research in semantic/pragmatic cartography theory. However, the relation of such map class entities to other entities in ontologies has been to date unclear, and even largely unconsidered. When taken seriously, however, the idea leads to a multi-level abstraction framework for ontologies—LEKXIS—that not only resonates with scientific knowledge and practice, but also helps to explain the occasional need in conceptual modeling to use concepts as instances. An outstanding issue in this respect is how concepts are pragmatically related between LEKXIS levels using the pragmatic square.

The resonance with scientific practice is also important. It not only eases knowledge capture but also knowledge use, because the knowledge itself is organized in ways that can readily be translated into common scientific activities. It should then not only lead to more interoperable ontologies, because their contents will use similar organizing principles concordant with geoscientific practice, but also to more interoperable tools and interfaces built around those principles. Reduction in semantic confusion, resulting from mitigation against the use of over-granular or under-granular concepts for some task, is another significant potential benefit of LEKXIS ontologies.

Apart from the implications for ontologies, the resonance with science also helps fill a theoretical vacuum, particularly in geography where situated concepts are especially intriguing because of their profound geographical orientation, i.e. their reliance on geospace-time as distinguishing criteria. The theoretical void they fill there lies between theories (that contain universally applicable concepts) and individuals (that are unique single entities), where they provide a conceptual bin in which to put generalized descriptions for a geographical case or situation, that are not universal but are local, such as those often made by human geographers.

Situated concepts and LEKXIS are, however, challenged by some very well known ontology engineering issues. These include known constraints on reasoning in second-order knowledge representations such as those in which concepts can be instances. Such constraints imply that, for second-order order logic statements, reasoning engines cannot: (1) guarantee to evaluate a statement in some reasonable number of steps (i.e. decidability), or (2) guarantee that all processable statements are valid (i.e. soundness), or (3) ensure that all logical implications of a statement can be determined (i.e. completeness). Strategies for coping with these restrictions need to be further explored in order to engineer viable
computing environments for LEKXIS. The outstanding question is whether the benefits to representation, grounded in the belief that there is a significant improved fit to geoscience knowledge, will outweigh the extra computational burden. Only additional prototyping and testing will ultimately tell. But, if recent computing trends can be used as an indicator, then the direction is clear: computational bottlenecks in hardware and software (including algorithms and reasoning) are being rapidly overcome, and attention is turning to hard knowledge-oriented problems, centered on ontologies for the moment, as generally envisioned by Newell more than twenty years ago (1982). Then getting the knowledge right represents today’s computing challenge, which in its own small way is what this dissertation is about.
REFERENCES


York, 152 pp.
Autonomous Database Systems. Morgan Kaufmann.
Emmence, C. (2002). The Chicken and the Orphean Egg: On the Function of Meaning and
381 pp.
discovery in databases. AI Magazine, Fall 1996, 37-54.
Feist, G. J. , Gorman, M. E. (1998). The psychology of science: review and integration of a
Hendler, J., Lieberman, H., Wahlster, W. (Eds.), Spinning the Semantic Web: Bringing
Web for Knowledge Management. http://www.aifb.uni-
karlsruhe.de/WBS/sst/Research/Publications/otk-bookv6.pdf
Flach, P., Kakas, A. (2000). Abduction and Induction: Essays on their Relation and
Integration, Kluwer.
sections with a chronology of geologic events. In: Proceedings of the 5th International
Symposium on Spatial Data Handling, IGU Commission on GIS, August 3-7, 1992,
Charleston, S.C., 544-553.
knowledge sharing in urban GIS. Computers, Environment, and Urban Systems
24(3):251-272.
Fonseca, F., Martin, J. (2004). Space and Time in Eco-Ontologies. AI Communications - The
European Journal on Artificial Intelligence, 17(4):259-269.
Ontologies: Information Engineering as a Hermeneutic Enterprise. Journal of the


This appendix describes the conversion of the database, as recorded by the geologists in the field, to a form acceptable for input to the neural network or to MMD clustering. Specifically, the multi-table relational database containing field data was de-normalized into one table, with each row representing the data for one observed site and with the columns possessing numeric values describing the site. Two stages of preparation were required: the nominal attribute values populating the attribute columns of the database, i.e. the attribute domains, required conversion to numeric values, and the relationships between data tables, i.e. the database schema, needed denormalization to one numeric vector per site.

The field database structure is depicted in Figure A.1, some example data for one site are shown in Figure A.2, and Table A.1 lists a representative set of attribute values for a subset of the attributes, including the hierarchical encoding for the rock type and structural feature attribute values. Note, of the five main tables populated with data only three, STATIon, LITHOlogy, and STRUCTure, were used as input to the geocomputational analysis; the remaining SAMPLE and PHOTOgraph tables either possessed information with limited direct contribution to concept development (i.e. samples) or could not be readily transformed for input to clustering (i.e. photographs). Transformation complexity also caused the free-format text comments in the three main tables to be omitted; digitally collected sketches were also omitted for similar reasons. Furthermore, the STATIon attributes were included in the input vector for information purposes only. Clustering thus occurred over the lithological and structural attributes, as suggested by geological theory, illustrating the impact of theory on the selection and preparation of significant attributes.

![Figure A.1: the field database schema.](image-url)
A.1 Attribute Preparation

A.1.1 Attribute Ranges
The attribute ranges consisted of numeric and nominal data types. Although some numeric attributes were ordinal (e.g. the UTMZ elevation value) others represented cyclic or geospatial data, requiring conversion to ordinal or ratio values for valid input to clustering. The remaining nominal attributes had ranges arranged as a partially ordered set (e.g. a rock type hierarchy), an unordered set (e.g. a list of minerals), or an ordered set (e.g. a list of deformation intensity values such strong, moderate, weak, etc.) Moreover, not only can attribute ranges have degrees of ordering, but the attribute instances can also be multi-valued and ordered: i.e. where an attribute instance possesses multiple values chosen from the range and where the ordering of the multiple values is significant, such as a mineral assemblage specified in order of decreasing abundance. In this study multi-valued attribute instances occurred as unordered sets drawn from unordered ranges, e.g. as lists of primary minerals or structural modifiers drawn from their respective domains. Other combinations such as unordered sets of values drawn from a partially ordered range were not encountered in the database, and are not considered herein.
<table>
<thead>
<tr>
<th>LITHO.ROCKTYPE</th>
<th>STRUC.FEATURE</th>
<th>LITHO.TEXMOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>20000 PLUTONIC</td>
<td>1000 PLANAR</td>
<td>acicular</td>
</tr>
<tr>
<td>20220 alkali feldspar granite</td>
<td>1100 BEDDING</td>
<td>acicular</td>
</tr>
<tr>
<td>20420 syenogranite</td>
<td>1110 subed</td>
<td>amygdaloidal</td>
</tr>
<tr>
<td>20460 monzogranite</td>
<td>1120 s-bed</td>
<td>aphanitic</td>
</tr>
<tr>
<td>20480 granodiorite</td>
<td>1130 sobed</td>
<td>aphyric</td>
</tr>
<tr>
<td>20620 tonalite</td>
<td>1200 PILLOW</td>
<td>autocratic</td>
</tr>
<tr>
<td>20800 syenite</td>
<td>1210 suplw</td>
<td>cumulate</td>
</tr>
<tr>
<td>20840 quartz syenite</td>
<td>1220 s-plw</td>
<td>cobble</td>
</tr>
<tr>
<td>21000 monzonite</td>
<td>1230 soplw</td>
<td>equigranular</td>
</tr>
<tr>
<td>21020 quartz monzonite</td>
<td>1300 IGNEOUS LAYERING</td>
<td>extrusive</td>
</tr>
<tr>
<td>...</td>
<td>1310 sulay</td>
<td>eyes</td>
</tr>
<tr>
<td>22000 gabbro</td>
<td>1320 s-lay</td>
<td>fibrous</td>
</tr>
<tr>
<td>22020 quartz gabbro</td>
<td>1330 solay</td>
<td>flow</td>
</tr>
<tr>
<td>...</td>
<td>1400 FLOW CONTACT</td>
<td>...</td>
</tr>
<tr>
<td>40000 HYPABYSSLAL</td>
<td>1410 suflow</td>
<td>glomeroporphyrític</td>
</tr>
<tr>
<td>40440 feldspar porphyry</td>
<td>1420 s-flow</td>
<td>homogenous</td>
</tr>
<tr>
<td>40600 diabase</td>
<td>1430 sofow</td>
<td>homogeneous</td>
</tr>
<tr>
<td>41000 lamprophyre</td>
<td>1500 FOLIATION</td>
<td>...</td>
</tr>
<tr>
<td>60000 VOLCANIC</td>
<td>1510 sufol</td>
<td>LITHO. METASSEM</td>
</tr>
<tr>
<td>60000 pyroxenite</td>
<td>1520 s-fol1</td>
<td>Hem</td>
</tr>
<tr>
<td>60200 felsic</td>
<td>1530 s-fol2</td>
<td>Mag</td>
</tr>
<tr>
<td>...</td>
<td>1600 FAULT</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>1700 SHEAR ZONE</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>1900 DYKE</td>
<td>...</td>
</tr>
<tr>
<td>80000 METAMORPHIC</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>82000 metasedimentary</td>
<td>2000 LINEAR</td>
<td>...</td>
</tr>
<tr>
<td>82200 Pelite</td>
<td>2100 EXTENSION LINEATION</td>
<td>Bt</td>
</tr>
<tr>
<td>...</td>
<td>2110 lu</td>
<td>...</td>
</tr>
<tr>
<td>82400 schist</td>
<td>2120 l-1</td>
<td>...</td>
</tr>
<tr>
<td>82600 psammite</td>
<td>2130 l-2</td>
<td>LITHO. ECOMINERAL</td>
</tr>
<tr>
<td>...</td>
<td>2200 MINERAL LINEATION</td>
<td>...</td>
</tr>
<tr>
<td>84000 metaigneous</td>
<td>2210 lumin</td>
<td>Py</td>
</tr>
<tr>
<td>84200 orthogneiss</td>
<td>2220 l-min1</td>
<td>Ccp</td>
</tr>
<tr>
<td>...</td>
<td>2230 l-min2</td>
<td>...</td>
</tr>
<tr>
<td>90000 structural metamorphic</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>90200 mylonite</td>
<td>2300 INTERSECTION LINEATION</td>
<td>boudinage</td>
</tr>
<tr>
<td>...</td>
<td>2400 FOLD AXIS</td>
<td>...</td>
</tr>
<tr>
<td>100000 SEDIMENTARY</td>
<td>2500 Z FOLD</td>
<td>...</td>
</tr>
<tr>
<td>100200 arenite</td>
<td>2500 fractured</td>
<td>...</td>
</tr>
<tr>
<td>100220 quartz arenite</td>
<td>2600 S FOLD</td>
<td>...</td>
</tr>
<tr>
<td>100600 conglomerate</td>
<td>2600 S FOLD</td>
<td>...</td>
</tr>
<tr>
<td>108000 siltstone</td>
<td>2700 U FOLD</td>
<td>...</td>
</tr>
<tr>
<td>102000 iron formation</td>
<td>2800 VEIN</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>2900 SLICKEN STRIAE</td>
<td>...</td>
</tr>
</tbody>
</table>
Partially Ordered Ranges

Both physical and social scientists may record information using fixed or evolving taxonomies that are hierarchically organized. The numeric representation of a taxonomy is reasonably straightforward, provided the hierarchy is not large or complex. In such cases, each branch of the hierarchy is assigned a range of digits for enumerating elements at that level. These numbers are concatenated along the branch, from parent to child, to form a hierarchic code that ensures that similar values are proximal. For example, the three level branch PLUTONIC→syenite→quartz syenite might be encoded, using one digit per level, as 100→140→142 and stored as the integers 100, 140, 142 respectively; a sibling branch such as PLUTONIC→monzonite→quartz monzonite could then be encoded as 100→150→151. Problems arise when hierarchies are wide or deep (i.e., have many elements at a level or many levels): for example, the encoding suggested for the hierarchy fragments above would fail if a level possessed more than nine entries and thus require two digits to encode the level; indeed, one digit was used for the first and third rock type levels and for the most populous second level two digits were used, encoding the ‘syenite’ fragment as 1000→1040→1042. In addition, the rocktypes in Table A.1 were actually scaled by a factor of 20, to result in 20000→20800→20840 for the syenite fragment (see discussion on scaling below in section A.1.3). Long or wide hierarchies, however, will result in codes that could exceed the computational limits of numeric representation which have fixed upper limits on the number of digits that can be represented for ratio numbers and ordinal values, e.g. the maximum integer represented on 64 bit processors is 2,147,483,647. More sophisticated numeric encoding of hierarchies face similar constraints (e.g. Teuhola, 1996).

Complex hierarchies are also problematic in that an element may be classified under more than one parent, reflecting diverse and possibly mixed clustering criteria and resulting in tangled trees. In this study, the few cases of multiple parentage were resolved pragmatically, by inspecting the data and noting that the context of the data involved one parent in almost every instance—the element was thus encoded as belonging to that parent. This resulted in straightforward hierarchies where conceptually similar terms were numerically proximal.

Thus, attributes with partially ordered domains, such as rock type (LITHO.ROCKTYPE) and structural feature (STRUC.FEATURE), were transformed into a single dimension in the input feature space and possessed a single partially ordered numeric value for each vector in the space.

Unordered Sets

An unordered set is a collection of data elements where the order of elements is not significant. Both attribute domains and attribute values can be unordered sets. For example, in this study the domain of minerals is treated as an unordered set from which the three mineral attributes, LITHO.METASSEM, LITHO.PRIMINERAL and LITHO.ECOMINERAL, can draw multiple unordered values, e.g. “Bt-Hbl”; structural modifiers in LITHO.STRUCMODE are treated similarly, e.g. “foliated, recrystallized”. Although the mineral domain can in fact be represented as a partial order, the resultant impact on clustering was thought to be minimal and even detrimental, in that the numeric values required to encode the mineral hierarchy would be large relative to the rock type codes and thus unduly weight the significance of the minerals, vis a vis rock types.
An attribute domain that is an unordered set can be encoded by projecting the domain elements into the input vector as individual binary attributes (Garson, 1998). Consider a vector \( v_1 \) consisting of \( k \) attributes \( a_i, 1 \leq i \leq k \), where the domain of the \( k \)th attribute \( m (a_k=m) \) consists of, for example, \( d_1 \) to \( d_n \) mineral names (e.g. \( d_1 = 'biotite', d_2 = 'hornblende' \), etc.), such that \( v_1 = (a_1, a_2, a_3, \ldots, a_{k-1}, m) \); then \( v_1 \) would be transformed to \( v_1* = (a_1, a_2, a_3, \ldots, a_{k-1}, d_1, d_2, d_3, \ldots, d_n) \) where \( d_i = \{0, 1\} \) to indicate absence or presence. Taxonomists (e.g., Clifford & Stephenson, 1975; Dunn & Everitt, 1982) indicate that multi-attribute binary encoding such as this lends itself to representation by a single number (e.g. \( m = '1 0 0' = 4 \)), obviating the need to project the domain into the attribute space. However, typical 32 or 64 bit numeric representation places an upper limit on the number of terms in the domain that is insufficient when \( n \) is large (e.g. greater than 32 or 64). Moreover, binary encoding is also inappropriate when the span of the \( d_i \) are extended (e.g. \( d_i = \{0,1,x\} \) (unknown)) or scaled (e.g. \( d_i = \{0,100\} \)), as was required by the study data (see Attribute Scaling below).

The scaled attribute domains of the study data required the domain values to be projected as dimensions into the feature space; i.e. each distinct domain value used, such as “aciccular”, “Bt”, “foliated”, became a distinct dimension in the input vector where its presence at a site was denoted by the value 100 and its absence by 0; unknown values were not recorded in the database and hence not encoded (though such encoding could occur via assignment of an arbitrary numeric value). For example, the attributes relevant to clustering in the LITHO schema in Figure A.2 are transformed from \( v_1 = (\text{ROCKTYPE, TEXMOD, METASSEM, PRIMENERAL, ECOMINERAL, STRUCMOD, ...}) \) to \( v_1* = (\text{ROCKTYPE, aciccular, ..., recrystallized, ..., hem, ..., bt, ..., pl, ..., qtz, ..., boudinage, ..., foliated, ..., gneissic, ...}) \), and the recorded observation in 10b then becomes the vector \((41300, 0, ..., 100, ..., 0, ..., 100, ..., 100, ..., 0, ..., 100, ..., 100, ...)\) when the ROCKTYPE attribute is scaled by a factor of 10.

Of note is that the projection of the attribute values into feature space dimensions can also accommodate multi-valued attribute instances that are unordered sets, in that absence/presence can be indicated for any number of domain elements in the input vector. A side effect of this technique is that it results in relatively empty feature space as a small number of the domain’s dimensions are populated with non-zero values. The sparse feature space was not deemed problematic for clustering insofar as all large feature spaces are relatively empty (Landgrebe, 1999) and because such sparseness has little impact on the SOM’s Euclidean distance similarity measure, which is applicable to both sparse or dense feature spaces.

**Ordered Sets**

Ordered sets are sequenced assemblages of data elements. Both attribute domains and attribute instances can be ordered sets.

Ordered attribute domains are easily represented as ordinal values. For example, the domain for STRUC.INTENSITY consists of values that range from strong to weak deformation intensity: e.g. intense = 70, very strong = 60, moderate strong = 50, strong = 40, moderate = 30, weak = 20, very weak = 10.

Ordered attribute instances whose domain is unordered can be handled by adopting the binary projection technique discussed above for unordered sets and replacing the binary encoding (i.e. \( d_i = \{0, 1\} \)) with an ordinal value (e.g. the sequential order, \( d_i = \{1,\ldots,n\} \)).
Although, this may introduce undesirable scaling effects when the range of $d_i$ is large, it acts to lessen the potential of semantic conflict when attributes are numerically discriminated. The extension of this projection method to interval data (e.g., % of abundance) or ratio data (e.g., quantity of abundance), with corresponding adjustments to the values of $d_i$, is analogous and straightforward.

It is clear that at most two pieces of information are captured by this method, element presence and some other concept such as sequence or abundance, and it is also clear that 3 or more orthogonal pieces of information are not accommodated (e.g. presence, significance (sequence) and abundance: ‘biotite(20%)-hornblende(20%)’), requiring the data to be normalized in the database sense (e.g. Date, 1990) and resulting in multiple occurrences within a sample:

<table>
<thead>
<tr>
<th>Station</th>
<th>Minerals</th>
<th>Station</th>
<th>Significance</th>
<th>Mineral</th>
<th>Abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site1</td>
<td>biotite(20%)-quartz(10%)</td>
<td>Site1</td>
<td>1</td>
<td>biotite</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Site1</td>
<td>2</td>
<td>quartz</td>
<td>20%</td>
</tr>
</tbody>
</table>

The transformation of multiple concept occurrences within a sample, and its application to the study data, is discussed below in the Schema Preparation Section.

**Cyclic Domains**

Cyclic domains are a special case of interval data where the scale is constant but cyclical, e.g., angles, days, weeks, months etc., where the beginning and end of the cycle are proximal such as Jan. and Dec. In the study data, the horizontal component of three dimensional orientation measurements (i.e. STRUC.AZIMUTH) was converted from an interval value (with 0-360 degree scale) to a Cartesian x-y polar coordinate with radius 100, to ensure that 0 and 360 are proximal; i.e. the attributes relevant to clustering in the STRUC schema were transformed from $v_1 = (\text{FEATURE, AZIMUTH, DIPLUNGE, INTENSITY, \ldots})$ to $v_1^* = (\text{FEATURE, AZIMUTH}_X, \text{AZIMUTH}_Y, \text{DIPLUNGE, INTENSITY, \ldots})$, and the recorded observation in Figure A.2 of (sufol, 315, 70, MOD-STR,\ldots) then becomes the vector (1510, -70.71, 70.71, 70, 50,\ldots).

**A.1.2 Spatial Data**

A distinctive trait of geographic thought is the notion of spatial correlation—the premise that spatial proximity endows thematic similarity. Spatial correlation may be induced in a SOM by encoding the site location in the input data vector to ensure actual nearest neighbors are proximal. The disadvantage of this approach lies squarely in its intent, as inappropriate scaling of the spatial attributes may cause them to dominate clustering, resulting in SOMs that reflect the geographic (versus thematic) distribution of the data. For example, sites that straddle a polygon boundary and are geographically proximal but thematically distant (i.e. belonging to different unit classes) could be evaluated as similar if the spatial data are incorporated and weighted heavily (Gahegan, 2000a). Consequently, the geospatial site coordinates were not included in the input vectors for the SOM, though they were clustered independently with MMD to reveal field data sampling patterns.
A.1.3 Attribute Scaling
SOM/LVQ calculations are sensitive to the scale and thus variance of the input data. They compare input vectors using a similarity measure, typically Euclidean distance, causing attributes with larger ranges and typically greater variance to dominate clustering. Performing statistical normalization on all attributes (by subtracting the mean and dividing by the standard deviation) will transform attributes to a common scale and temper overt dominance (Garson, 1998).

Alternatively, weighting functions may be applied to specific input variables, to reflect their relative value to clustering. Selecting appropriate weighting factors is rather arbitrary (Garson, 1998; Kohonen, 1995) and exceeds purely inductive, data-driven, approaches, and encroaches on theory-based methods, where criteria beyond the raw data exert influence on clustering. Such criteria may consist of the laws of science, explanatory models and their exemplar occurrences, as well as rules of thumb and tacit knowledge accumulated via general experience and from prolonged interaction with a specific study area. As geological mapping involves the construction of a 3d spatio-temporal model for an area, some attributes possess greater concept development relevance than others, while the others contribute to the geometric, topologic or age description of the evolving individual model. The degree of conceptualization relevance is determined both a priori, in a general theoretic sense from the model-based criteria, and in situ, where local factors may refine or supplement theory. Experimentation with the SOM/LVQ demonstrated that scaling the rock type attribute by a factor of 20 to reflect their significance within the local model and theory was one of two main factors leading to adequate clustering; the other being successful conversion of the database schema.

A.2 Schema Preparation
Once attribute values were converted to numeric values, then tables containing these values were denormalized to a single table, in first normal form (Date, 1990), in which each row represented a site vector suitable for input to the neural network or MMD. Site vectors are desirable because the SOM/LVQ is sensitive to both the distribution and frequency of data, hence sites with more observations would exude greater influence on a map unit concept; conversely, providing representative site vectors ensures each site is weighted equally. Denormalization is problematic when field sites are described by multiple occurrences, such as multiple lithologies or structural measurements, and when these occurrences are themselves inter-related, such as when many structural measurements are taken within a specific host lithology. Both these issues were encountered in the field database schema: a one-to-many cardinality existed between sites and lithologies, sites and structural measurements, and between lithologies and the structural measurements (Figure A.2, above). These relationships suggest a denormalization strategy in which representative structural measurement vectors are first developed for each lithology, and then representative lithology-structure vectors are developed for each site.

Two methods of determining representative vectors were used: (1) selection directly from data, and (2) development of a single representative vector from multiple site data using the SOM. For lithologies this entailed, respectively: (1) selection of all the dominant lithologies at a site (indicated by LITHO.ROCKNUM = 1), and (2) development of a single representative vector for all site lithologies. For structural measurements this entailed, respectively: (1) selection of the planar structural measurement type (e.g. foliation) or...
selection of the full planar structural measurement (e.g. foliation, azimuth, dip,…) , and (2) development of a single representative vector from all relevant structural measurements, using the SOM. Six sets of denormalized site vectors were then prepared using these approaches; the sets consisted of vectors for: one or more dominant lithologies per site, one or more planar structural measurement type per dominant lithology per site, one or more full planar structural measurement per dominant lithology per site, dominant site lithologies plus a SOM-derived representative structural measurement vector, a SOM-derived representative vector for all lithologies at a site, and a SOM-derived representative vector for all site lithologies where each lithology includes a SOM-derived representative structural measurement vector. The objective behind this partitioning of data and attributes was to identify which data and attributes contributed to clustering, based on the intuition that some data and attributes contribute in large part to concept discovery and should lead to increased clustering, while others contribute instead to classification, such as the development of polygon boundaries on the map, and/or to the description local complexities disjoint with the concept, such as the presence of anomalous lithologies within a map unit, leading to increased confusion.

SOM-derived representative vectors were developed by using the SOM as a hashing function to compress multiple input vectors into a single representative output vector. This involved mapping a set \( A \) of \( n \) input vectors of length \( m \) to a single output vector \( S \) of length \( k \) using a SOM configured as a \( (i,j) \) matrix with \( r \) rows and \( c \) columns, such that \( 0 \leq i < r \), \( 0 \leq j < c \) and \( r \times c = k \). A representative vector for the \( n \) inputs was then determined by selecting for each input the nearest \( i,j \)th SOM vector, via Euclidean distance, scaling the responding distance (i.e. quantized error) between the input and responding SOM vector and placing it into the output vector at position \( p \), \( 0 \leq p < k \), determined via row order (\( p = ci + j \)) (Figure A.3); column order (\( p = rj + i \)) could also have been used.

Scaling involved statistical standardization of the responding distance. To ensure that dominant lithologies exuded greater influence on clustering, the responding distance was then additionally scaled by a factor of 10 for the dominant lithologies and reduced by the same factor for the remaining lithologies. Several collisions occurred at \( p \) when input vectors converged to the same \( i,j \)th SOM position as a result of conceptually similar data, i.e. when...
many similar data were recorded at a site. Collisions were resolved by: (1) selecting a dominant lithology during site vector development, otherwise (2) accepting the last colliding value at $p$. 
BOYAN BRODARIC

EDUCATION
1999-2005   The Pennsylvania State University   Ph.D., Geography, 2005
1981-1988   The University of Toronto   B.Sc., Math / Philosophy, 1988

EMPLOYMENT
1990-present   The Geological Survey of Canada   Physical Scientist
1986-1990   The Ontario Geological Survey   Information Scientist

SELECTED PUBLICATIONS