Re-Envisioning Data Description Using Peirce's Pragmatics

Mark Gahegan and Benjamin Adams

Centre for eResearch, The University of Auckland, New Zealand

Abstract. Given the growth in geographical data production, and the various mandates to make sharing of data a priority, there is a pressing need to facilitate the appropriate uptake and reuse of geographical data. However, describing the meaning and quality of data and thus finding data to fit a specific need remain as open problems, despite much research on these themes over many years. We have strong metadata standards for describing facts about data, and ontologies to describe semantic relationships among data, but these do not yet provide a viable basis on which to describe and share data reliably. We contend that one reason for this is the highly contextual and situated nature of geographic data, something that current models do not capture well — and yet they could. We show in this paper that a reconceptualization of geographical information in terms of Peirce's Pragmatics (specifically firstness, secondness and thirdness) can provide the necessary modeling power for representing situations of data use and data production, and for recognizing that we do not all see and understand in the same way. This in turn provides additional dimensions by which intentions and purpose can be brought into the representation of geographical data. Doing so does not solve all problems related to sharing meaning, but it gives us more to work with. Practically speaking, enlarging the focus from data model descriptions to descriptions of the pragmatics of the data — community, task, and domain semantics — allows us to describe the **how**, **who**, and **why** of data. These pragmatics offer a mechanism to differentiate between the perceived meanings of data as seen by different users, specifically in our examples herein between producers and consumers. Formally, we propose a generative graphical model for geographic data production through pragmatic description spaces and a pragmatic data description relation. As a simple demonstration of viability, we also show how this model can be used to **learn** knowledge about the community, the tasks undertaken, and even domain categories, from text descriptions of data and use-cases that are currently available. We show that the knowledge we gain can be used to improve our ability to find fit-for-purpose data.

1 Rethinking the Way we Describe Geographic Data

Our efforts to create better geographic data models and communicate richer data descriptions have led to very fruitful avenues of research, such as the representation of semantics, the visualization of uncertainty, the propagation of error, and others [43,18,41,26,21,28]. The era of volunteered geographic information (VGI) further complicates the picture with new challenges for understanding spatial data meaning, accuracy, and quality [19,1]. Research to date may allow us to describe the quality

M. Duckham et al. (Eds.): GIScience 2014, LNCS 8728, pp. 142–158, 2014.

(or perhaps even the semantics) of a single dataset, with effort, but we cannot propagate — with suitable modification — this information into derived products. Thus the onus remains firmly on the data producer to document quality and meaning of every new dataset. This has never been sustainable; most datasets do not have comprehensive quality information at the level of sophistication that consumers need. It is even less sustainable in the era of VGI and mash-ups, where more data is combined in hitherto unanticipated ways than ever before.

Furthermore, there is a real danger that all these different research strands have moved us further away from the actual problem, of describing these important aspects of data in an integrated and combinable manner, for example so that they can be used together in a query to find useful data. Without a way to bring these threads back together, our fruitful research avenues are in danger of becoming *cul-de-sacs*. Our approach to modeling geographic data is drastically in need of an overhaul.

Finally, as a community, we have been guilty of concentrating too heavily on the perspective of the data producer: describing "facts" about data, but not acknowledging the tacit world-view that can render these "facts" true and useful (or not) within a given context. Knowing which "facts" remain true when the context is changed, and also which facts remain relevant are both key to describing geographical data better. We term this idea the *pragmatics* of data, after Peirce [32].

1.1 An Alternative Approach

We suggest the following five propositions offer an alternative way forward:

1. We do not know what the eventual user will need to know about the data they wish to use, and we cannot know, in advance, the likely utility of any of the descriptions we may strive to add as data producers (such as ontologies, workflows, and accuracy assessments). And despite the huge volume of work published on conceptual geographic data models[1], we are no closer to knowing which ones have lasting value. We need empirical evidence, not more rhetoric, to produce a better model.

2. Consequently, we deliberately move away from the search for a single, definitive conceptual model of geographical data, and propose instead a meta-model where we can evaluate the actual utility of various forms of descriptions, from the perspective of specific tasks and research needs, using evidence gathered from actual use-cases.

3. We propose this simple meta-model as a set of description spaces, each comprised as "facets," that represent themes that we believe may have utility — but we do not claim that these are either necessary or sufficient — they are rather a place to begin. Within these facets we measure compliance to some kind of desired "optimal" state — as simply as we can (see section 2). Again, we make no claim that these facets are right, rather that they may prove to be useful under evaluation and (hopefully) that they are simple enough to be assigned and read with ease.

4. We broaden the scope of data description to consider the perspective of the data consumer. So we begin by asking: "What kinds of things might a consumer of the data want to know?" Rather than: "What kinds of things might a producer of the data be

[1] Including work published by the authors of this paper!

persuaded to say?" Furthermore, current approaches emphasize the **where**, **when**, and **what** aspects of data, with various degrees of success and completeness, but often leave aside the deeper questions of **who**, **how**, and **why**. These questions carry much meaning for a potential consumer of the information (they speak to reputation, quality, and motivation). We believe there are aspects of these deeper questions that can be captured that allow us to start framing more practical (and answerable) questions that often substitute for deeper ontological and epistemological questions: e.g., "for what task did you make this data?" can act as a surrogate for: "what does this data mean to you?" or "which organization produced the data?' may in some circumstances substitute for "what is the likely quality of the data?" These substitutions are certainly not perfect, but in a Bayesian sense they are better than nothing; and what's more, we can readily compute the degree to which they help elucidate the pragmatics we seek, as we show in Section 4.

5. The benefits of such an approach are many: (i) descriptive facets can be added or retracted according to need; (ii) the system could learn over time which kinds of data descriptions are most useful, so that data producers can focus their efforts when creating time-consuming data descriptions; (iii) multiple perspectives onto the meaning and use of data can be supported concurrently — allowing for the natural fact that we do not all see the world in the same way; (iv) shifting the emphasis from producing more metadata to learning from use-cases lifts an unmanageable burden from the data producers; (v) the conceptual model is not now a fixed thing, but can grow or change as new needs arise, as we learn more about which facets offer the most useful descriptions of data, or as new computational technologies provide us with additional descriptive facets.

The following are some of the many important facets to describe, though of course not an exhaustive list:

- Data Model: **What/when/where** is it?
  - Spatio-temporal Frameworks (spatio-temporal schema & semantics)
  - Attribute Schema & Semantics
- Process: **How** was it made and thus how confident are we in it?
  - Quality (Accuracy & Uncertainty)
  - Provenance (lineage)
- Community: **Who** can/should use it? **Why** was it made?
  - Motivation
  - Access and licensing
  - Authority (Governance & Trustworthiness)

1.2 Background

The description of geographic data into distinct spatial, temporal, and thematic components (**where**, **when**, and **what**) pre-dates modern geographic information systems and goes back at least to Berry's geographic matrix [3]. This matrix has formed the basis for much of the conceptual modeling surrounding geographic data into logical systems for representing geographic units [17]. Conceptual modeling in GIScience has looked at many dimensions of geographic and spatial information, including the object/field distinction, spatial relations, temporal relations [27,33,45]. Representation of the semantics of attributes using object-oriented databases and formal semantics continues to

be an active area of GIScience research [11,10,24]. However, by ignoring **how**, **who**, and **why** these models (explicitly or implicitly) take either an exclusive producer's view on what the data means, or attempt to describe a universal view; in either case without situatedness, or context. When context has been studied it has been operationalized in terms of weights on attributes for semantic similarity measurement — not in terms of process and community [36,39,23]. But we need this situatedness to allow us to differentiate between the perspectives that naturally arise with a community, for example between of the producer of the data and the eventual consumers, particularly the unexpected consumers [14].

Philosophical Foundations: Peirce's Firstness, Secondness, and Thirdness. The representation of the situatedness of information is a natural consequence of acknowledging that we do not all see things the same, or that meaning and utility can depend on the situation at hand. C.S. Peirce [32] first proposed such a model, broadly based on semiotics, to demonstrate how signs are created and interpreted in communication.

> Peirce's notion of sign was broad enough to include situations, contexts, propositions ... and their expression in any language, including English and logic. His notion of ground is crucial: it acknowledges that some agent's purpose, intention, or "conception" is essential for determining the scope of a situation or context.[44]

In Peirce's pragmatics, **firstness** refers to a concept that remains constant when viewed from different points of view; it simply "is," and requires no qualification. An example might be the fact that a city's population is 1.5 million people. Firstness could also include the thematic aspects of the data as articulated in the attribute fields of the data. Similarity of features based on geometry and much of the semantic similarity measurement work done in geosemantics falls under this category [38,39,23]. Databases and GISystems are well equipped for representing this kind of information.

Secondness. refers to concepts that require further description or explanation via first-order relations to other concepts, but without the need for further interpretation or qualification. For geographic data this means the relationships to scientific conceptual knowledge that informs the data, such as: the tasks and scientific processes that consume or produce the data or the semantic commitments of the domain knowledge. For example, a city is a kind of settlement, or a city is bigger than a town. Similarity based on secondness is the similarity of tasks and domain knowledge during acts of production and use of the data. Ontologies and workflows represent this kind of information well.

Thirdness. adds a qualifier: two things are brought into relation only within the context of a third (i.e., relations of relations). In the case of data, thirdness can, for example, represent the community of people who accept as true a certain set of attribute values and semantic commitments (statements of firstness and secondness, respectively). For example, a concept that only has relevance or acceptance among a specific group, such as *provisioning services*, which is a notion accepted by scientists studying ecosystem services but not widely accepted by other ecologists. Thirdness forms the basis of pragmatic reasoning, that data and relationships may not be true in all circumstances or to all

participants, but may require interpretation in the light of experience or within a given situation. Thirdness measures of similarity are almost always overlooked but can prove valuable if one wants to find data that match community constraints. For example, data that fulfills a community's usage or personalizing data search based on matching user profiles [8,6].

Note that firstness, secondness, and thirdness are not necessarily fixed; at some time we may wish to assert a "fact," at other times we may challenge the same fact and wish to explore its foundation, or decide that it only applies within a specific context. Importantly, the nexus of interactions between community, scientific knowledge, and data can be examined from different perspectives [15,34]. Here we have focused on data as the immediate subject and looked at relations to that data. If we had taken the data producer as the subject, then firstness similarity would refer to qualities of the producer and the data could be modeled via thirdness relations that provide insight to the likely domain expertise of the producers.

2 Four Description Spaces: Data, Domain, Task, and Community

Here we propose a simple model for the pragmatic description of data that moves from *community* to *task* to *scientific domain knowledge* to *data description*, or visa versa. These four aspects of the pragmatics of geographic data provide a more complete context for understanding the meaning of data, or the fitness of data for various purposes, because they describe knowledge of community and knowledge of the underlying science along with the semantics and schemata of the data.

Using Peirce's categories as a guide we present a theory for comparing the similarity of geographic data based on firstness, secondness, and thirdness measures over these description spaces. We conceive of these spaces as having similarity metrics because that will allow us to define aspects of community, domain knowledge, task, and data as compact regions in the spaces. The similarity metric for the space can be defined in terms of categorical or set-theoretic similarity over a knowledge graph, such as a description logic ontology, or any of many other similarity strategies [13].

The important point is that each of these spaces has a number of facets that allow us to reason about the similarity of the instances in those spaces. The facets provide constraints by which we can match queries for data from a data consumer to the data objects that have been created by a producer in a potentially very different context. The facets can be as simple or as complex as needed, experience suggests that simpler is better because some descriptive information is better than none (because it is too demanding to supply). In our example below, simple ordinal statements implying a greater level of compliance to some agreed set of information goals might be a practical and useful approach for many tasks, although other approaches may be equally valid [7]. Table 1 lists two sample dimensions or facets for each of the four spaces we have described, which can be used to make compliance judgments for data to determine for example if it is fit-for-purpose. Each statement represents a progressively deeper commitment towards some ideal, and subsumes the previous commitments.

Figure 1 illustrates how two datasets can be represented across these spaces. The first dataset, represented by **o**, is historical monitoring data about water wells and aquifers

Table 1. This table shows eight ordinal facets that can be used to reason about compliance of data based on four pragmatic spaces

Community
Data Standing 0. No information 1. Intent behind the data is known (implies an understanding of the purpose beyond some threshold) 2. Data originates from a reputable source (implies community aspects are known beyond some threshold) 3. Peer review and repeated use has verified utility and quality of the data 4. Authoritative data source endorsed by community
Data Licensing and Openness 0. No information 1. Author publishes a link to the data 2. Data license and reuse terms are known and published with the data 3. Data is available via persistent URI 4. Data is registered with an open SDI or similar cataloging service
Task
Process / Workflow 0. No information 1. Some aspects of the task can be inferred from knowledge of the community (and/or the data) 2. A clear description of the task is provided as text 3. A formal description of the task is provided (such as via a task or application ontology) 4. A full, repeatable workflow and associated data are provided, that allow the task outcomes to be repeated
Intention 0. No information 1. Some aspects of the intent can be inferred from knowledge of the community (and/or the data) 2. Clear text statement of intent or scientific goals behind the task 3. Description of intention using a controlled vocabulary 4. Detailed description of meta-level science model
Domain Semantics
Formality of domain semantics 0. No information 1. Informal concept maps of domain are provided 2. Controlled vocabularies used to describe data 3. Lightweight (Web) Semantic schema and SPARQL end points provided 4. Uses appropriate domain ontologies to describe semantics
Completeness of domain semantics 0. No information 1. Upper-level domain ontology for broad concepts (such as SWEET) [35] 2. Anchored into top-level ontology (such as Dolce) [16] 3. Detailed domain ontology (such as GeoSciML) [40]
Data Syntax and Attributional Semantics
Data Schema 0. No information 1. Spatial data correctly geo-registers (we know the projection, coordinate system, etc.) 2. Attribute schema is published and correct (we can actually parse the data content!) 3. Data is published using relevant (open) standards
Metadata (beyond data schema) 0. No information 1. A minimal metadata standard is met 2. Full metadata is provided using relevant open standards. 3. Validated account of data collection and interpretation process is available (such as a geological field manual for a mapsheet)

made available as part of the National Groundwater database (NGWD) by Natural Resources Canada's Earth Sciences Sector Groundwater program.[2] In the community space this dataset is at the higher end of both dimensions as it is an authoritative source and it is registered and made available on an official website. In the task space it scores a 1 on the process/workflow dimension as some aspects of the data collection process can be inferred from the data. It scores a 2 on the intention dimension because there are clear descriptions of important uses of the data on the Environment Canada website. Along the domain semantics dimensions it scores highly, because the concepts are de-

[2] `http://ngwd-bdnes.cits.nrcan.gc.ca/service/api_ngwds:def/en/presentation.html`

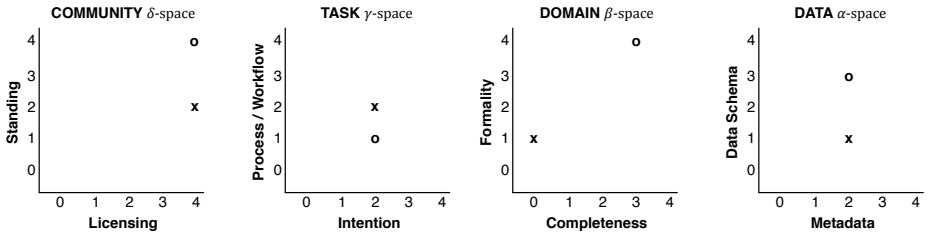


Fig. 1. Four domains of pragmatic and semantic description with example ordinal dimensions for reasoning about compliance. The **o** represents a dataset from NGWD and the **x** represents a dataset collected by the LTER network.

scribed using the GroundWater Markup Language (GWML) specification, an extension to GeoSciML [5]. Finally, it scores a 3 in the data schema dimension, because the data is easily parsed and linked to OGC open data standards, and it scores a 2 on the metadata dimension because the data attributes are fully described in the metadata.

The second example, represented by **x**, is a sample dataset of temperature and snow density data collected by a member of the Long Term Ecological Research (LTER) network [22]. The data originates from a reputable source (LTER), so scores 2 on data standing and is made freely available on the DataONE data network with DOI (`knb-lter-nwt.34.8`), so scores a 4 on the data licensing and openness [30]. It scores a 2 on both process/workflow and intention dimensions because the task and scientific goals are both clearly presented in the abstract associated with the dataset. It scores a 1 on the formality of domain semantics dimension because it is aligned with the LTER controlled vocabulary, but scores a 0 on completeness of domain semantics as that controlled vocabulary is described in SKOS, not a formal ontology. The geographic data schema correctly registers to WGS 84 coordinate system, so scores 1 on the data schema dimension. The attribute metadata dimension scores 2, because the metadata is described using the Ecological Metadata Language (EML) [12].

3 Generative Model for Geographic Data Creation

One goal of describing a model of geographic data semantics and pragmatics is to provide a mechanism to find data that are *fit-for-purpose*. The examples that follow assume this goal. From our point of view, we approach this goal by creating a representation of the communication act (both intentional and by implication) that is occurring between the producers and consumers of the data. This can be modeled using a graphical representation, as sets of relations. The generative process is illustrated in Figure 2 and demonstrates how a consumer and producer are indirectly linked to a data object via the description spaces of task and domain. This generative model is a sub-graph of the broader nexus of interactions that contribute to geographic (and other types of) understanding (see [14] for a more detailed discussion of this). Other models derived from this nexus are certainly possible, e.g., one might specify a direct edge between the community and domain space.

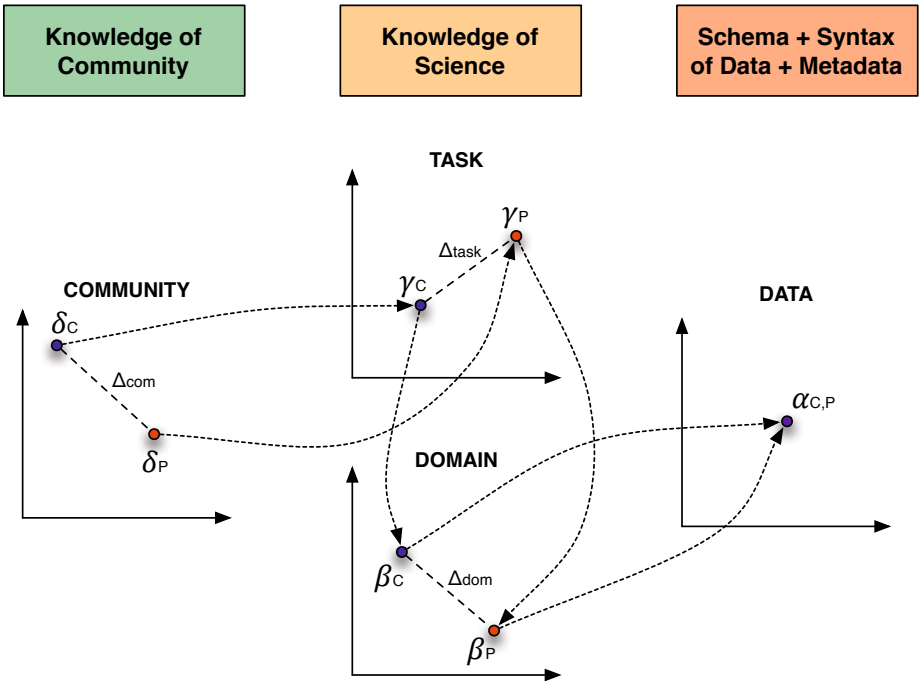


Fig. 2. Graphical model of data production and use via tasks and domain-specific semantic commitments. Consumer δ_C uses data object $\alpha_{C,P}$ with semantics β_C for task γ_C, although it was made by producer δ_P with semantic commitments γ_P for task β_P.

Data Model Space. Data in geographic information systems are often described in terms of geometry, other attributional characteristics, and occasionally temporal aspects. The data description space consists of dimensions that differentiate data along these respects. The GIS operations that transform data, e.g., projection and cartographic generalization, have the effect of moving a data instance from one point in this space to another. We can compare the similarity of two data objects based on their data description and this subsumes both traditional geometric matching, such as used in conflation algorithms, as well as similarity based on attribute value statistics. In our model all of these measures of similarity constitute *firstness* measures of similarity. They are based on characteristics of the geographic data artifacts themselves, divorced from interpretive modifiers. Within GIScience firstness measures of similarity have dominated the literature.

Domain Space. The domain-specific semantic commitments describe the semantics of the data in terms of a scientific domain. The interpretation of domain semantics can be restricted through the use of formal ontology, although the facets of this space do not necessarily need to be defined in this manner [20]. The work in geosemantics that looks at comparing the similarity of geographic concepts falls in this space and is a kind of

secondness similarity. It remains an open question whether practical merging of domain ontologies and concept similarity measurement across multiple ontologies is solvable, thus we deem it important to not only consider these semantics but look at the tasks for which the data is intended [42].

Task Space. It may be well that the tasks that one wants to perform with the data is a better indicator of fitness-for-purpose than similarity measurement based on the data description. For example, if a user wishes to model wildfire, and knows that a specific vegetation coverage was created for exactly this purpose, it may well be useful to explore it further and, if necessary, adapt their own methodology or conceptual understanding to use it. It is also perhaps more likely that such a coverage will use data models and make ontological commitments that will be in keeping with those of the user: a vegetation coverage created to explore species diversity may not be so suitable. Note that this claim is not necessarily true for any specific example datasets. There will undoubtedly be counter-examples, but the principle applies in the sense of increased likelihood.

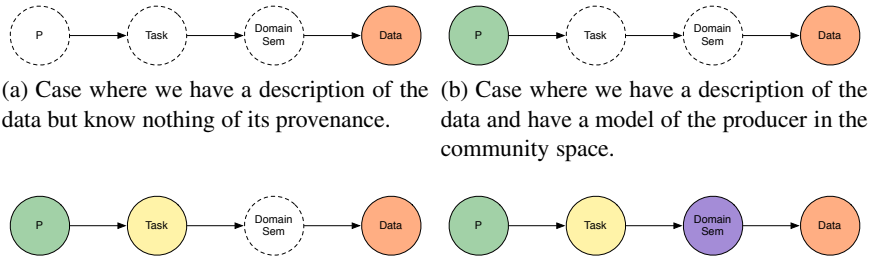
Community Space. The dimensions of the community space provide a means to describe the properties of both consumers and producers of data. Within this space we might recognize key themes and specializations that occur in the work of individuals and groups, constraints on information licensing and sharing, and governance issues about the authoritativeness of data. Based on usage, we may also be able to infer qualities such as trust and expertise [2].

3.1 What Variables do we Observe in the Graphical Model?

What we know about the pragmatics of geographic data will vary greatly from one data object to another. For example, it is possible that we might know nothing of the provenance of the data; we might only know the schema and attribute semantics of the data themselves. In other cases we might have information about the community, based on keeping track of use-cases, but no semantics or schema published. The model asserts that even when we do not have a full set of information available in relation to the data, the four description spaces can act as latent variables. For example, Figure 3a shows the case where we only have information about the data. Figures 3b–3d show cases where we know progressively more about the pragmatics of the data until we have a full picture with description of the producer in the community space, a description of the task in the task space, a description of the semantic commitments in the domain space, and a rich description of the data in the data model space. We explore later the question: To what extent can knowledge from one space provide insights into another?

3.2 Formalization

Formally, we define a *pragmatic data description* as a 4-tuple $< \delta, \gamma, \beta, \alpha >$, where δ is the community descriptor, γ is the task descriptor, β is the domain semantics descriptor, and α is the descriptor of data schema, spatio-temporal properties, and attribute semantics. The *context relation* $\boxdot$ is a binary relation between a symbolic data



(a) Case where we have a description of the data but know nothing of its provenance.

(b) Case where we have a description of the data and have a model of the producer in the community space.

(c) Case where we have a description of the data, have a model of the producer, and know the task for which the data is created.

(d) Ideal case where we have descriptions of the data, producer, task, and the semantic commitments from the scientific domain.

Fig. 3. Different degrees of observed pragmatics

object d (i.e., the actual digital encoding of the data) and its pragmatic data description: $\boxdot(d, <\delta, \gamma, \beta, \alpha>)$.

For simplicity's sake if we consider these spaces to be independent, then we can define fitness-for-purpose (ffp) as a compound distance measure across the firstness, secondness, and thirdness similarity spaces (Equation 1). In Section 4 we will discuss a probabilistic approach to measure the relatedness *between* the elements of the δ, γ, β, and α spaces.

$$ffp(i,j) = \Delta(\alpha_i, \alpha_j) + \Delta(\beta_i, \beta_j) + \Delta(\gamma_i, \gamma_j) + \Delta(\delta_i, \delta_j) \tag{1}$$

3.3 Consumer and Producer

This generative model starts at the producer and ends with data. We can use this model for the consumer as well if we swap the consumer in for the producer. We consider the problem of finding data that is fit-for-purpose as one of finding the ideal data object d^* given that we know the consumer's description within the community space, we know what task they want to perform with the data, and we understand the semantic commitments they have made. Thus, we want to find a data object d from the set of all data objects D that minimizes $ffp(d^*, d)$. (Practically speaking we want to find several

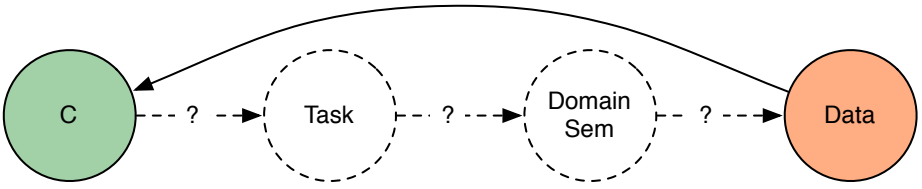


Fig. 4. The consumer wants fit-for-purpose data but the task and domain semantics are not-observable (latent variables in the generative model)

examples of data objects ordered by ffp, thus giving the user a few options to choose from.)

In many cases the task, the domain semantics, and even the data description will not be explicitly defined by the consumer and therefore these must be treated as latent variables in the model (see Figure 4). Realistically, we might be restricted to thirdness similarity measures in this case. For example, we might offer data from producers with a similar user profile.

4 Bayesian Interpretation for Learning and Prediction

The graphical model presented in the previous section points to mechanistic approaches to learn categories of geographic data by community, task, and domain semantics, in addition to traditional geosemantics. This is done by interpreting the graphical model described in the previous section as a Bayesian network, which provides significant statistical inferential power. A Bayesian network is a directed acyclic graph where nodes represent random variables and the edges represent their conditional dependencies. The directed edges between nodes are assigned probabilities and it satisfies the local Markov property that the variables are conditionally independent of other variables that are not parents in the graph [37]. The relationships between variables in a Bayesian network are often interpreted as causal relationships and can be used to model generative processes [31,4]. Thus, e.g., we can describe the probability that a producer δ_i will perform task γ_j. That task γ_j will entail domain semantics β_k and so on. Figure 5 shows an example of the Bayesian network that extends from a given producer and describes the probabilities of dependent tasks, domain semantics, and data descriptions.

Given a hypothesis space $\mathcal{H}$, we can use Bayes theorem to identify the most probable hypothesis, h, in that space to explain observed data, $\mathbf{d}$. $P(h)$ is the prior probability that a hypothesis is correct based on background knowledge. $P(\mathbf{d})$ is the prior probability that the data $\mathbf{d}$ is observed and $P(\mathbf{d}|h)$ is the probability that $\mathbf{d}$ is observed given that h is true. $P(h|\mathbf{d})$ is the posterior probability of h, i.e., what is the probability that the hypothesis holds given that $\mathbf{d}$ has been observed. We calculate this posterior probability by rewriting the denominator of Bayes' Rule (Equation 2).

$$P(h|\mathbf{d}) = \frac{P(\mathbf{d}|h)P(h)}{\sum_{h' \in \mathcal{H}} P(\mathbf{d}|h')P(h')} \tag{2}$$

By parameterizing the dimensions in the community, scientific knowledge, and data description spaces and maximizing posterior probability given a set of training data, it opens the possibility of induction of new classification and prediction methods based on firstness, secondness, and thirdness categories (and compositions of all three). The effectiveness of Naïve Bayes classifiers and other hierarchical Bayesian networks with latent variables are well established and can be directly applied to this model [9]. The challenge moving ahead is articulating the dimensions of these spaces such that we can use these machine learning methods.

A probabilistic generative model gives us a way to describe the *potential* kinds of geographic data that a source can generate. Based on the probabilities in the graph we can

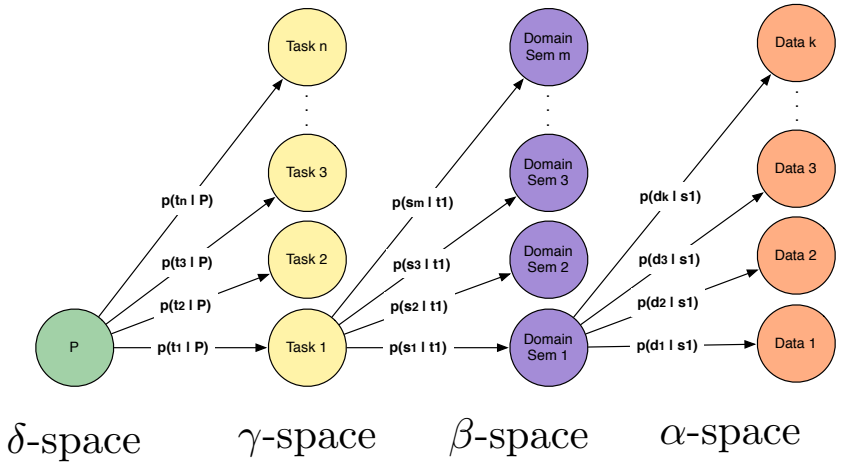


Fig. 5. Data production graph as a Bayesian network

see the data that are likely to be generated by an individual source represented in community space. In addition to unsupervised learning of categories of tasks, communities, and semantic commitments, we can ask questions about latent variables in the model given some other knowledge that is available. For example, (a) probability of data given task, given domain semantic commitments and given producer; (b) probability of task given data; and (c) probability of community category given task.

Figure 6 provides a schematic of the kind of results we can anticipate given descriptions using facets such as those described in Table 1, which defines pragmatic description in terms of four two-dimensional spaces. This figure shows that a measure of 2 along both the Process/Workflow and Intention dimensions in task (β) space probabilistically implies certain values in other dimensions. Different descriptions of pragmatics will lead to different results. Importantly, this gives us the ability to experiment with different kinds of data description, changing facets and their dimensions and generative relationships to compute their utility via information gain measures [29].

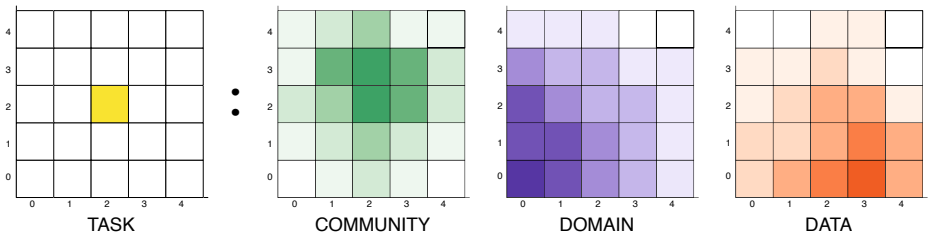


Fig. 6. Pixel-oriented visualization of probabilities of descriptions in the other three spaces, given a fixed point in Task space [25]. Darker colors represent higher probabilities.

4.1 Using Existing Descriptions

In order to perform this kind of unsupervised learning of categories of community, task, and domains; sufficient training data is required. But of course, much of the information we seek is not recorded directly in any current system. Large scale cyberinfrastructure projects like DataONE — a federated data network designed to enable discovery of environmental data — are beginning to address the problem of pragmatics [30]. Metadata describing purpose, method, authorship, rights holders, usage rights, and general abstracts written by the producers provide views to the pragmatics of the data, albeit often in unstructured natural language. A majority of the metadata that exist within DataONE have a geospatial component, but formal description of geosemantics and the geographic data model are virtually non-existent. In contrast, spatial data infrastructures are moving toward richer descriptions of geosemantics but broader pragmatics are largely lacking [24]. With some work, it is possible to assemble a rich enough description from which to begin.

As proof-of-concept of the network model, we downloaded 59,879 metadata descriptions from DataONE data objects that include geographic data. Although, the metadata do not describe pragmatics in the rich way we advocate earlier in this paper, we can demonstrate that by string matching terms that we associate with community members, methods and tasks, scientific domain knowledge and geographic representation we can find statistical pragmatic relationships. Figure 7 shows a small set of terms from this DataONE metadata mapped into a simple Bayesian network like the one shown in Figure 5. Since the DataONE metadata does not clearly differentiate between task and domain, we describe a simplified **science** description space, roughly covering both of the task and domain spaces as defined earlier in the paper. Once the network is built we use Markov chain Monte Carlo inference to find the likelihood of data given pragmatic evidence.

For example, we find that the probability of **fire**-related data given an **ecologist** producer is 8.0%, but when we add that the domain is **disturbance**, then the probability increases to 29.5%. Likewise, given a **climate scientist** producer there is only a 2.2% likelihood of precipitation data, but when the condition of **vegetation dynamics** is added then it rises to 58.5%. When scientific concepts are researched together, then it can imply high likelihood of data. For example, the probability of **tree** data given both **vegetation dynamics** and **disturbance** is quite high: 55.5%. The code and data for running these and similar experiments are available for download at `https://wiki.auckland.ac.nz/x/mBKsAw`.

To illustrate that these techniques can also be used to describe relationships between types of producers and data formats, Figure 8 shows how (in the DataONE network) a data format can be indicative of being useful for a specific community. For example, **hydrologists** are much more likely to do research with digital elevation model data (presumably due to their interest in catchment areas) than are **climatologists**. Whereas a NetCDF format strongly indicates relevance for a **climatologist**. Thus, a spatial data infrastructure that has user profiles of data consumers can provide a personalized data search service based on these results — e.g., suggesting DEMs if the system knows that the consumer is a hydrologist and so on.

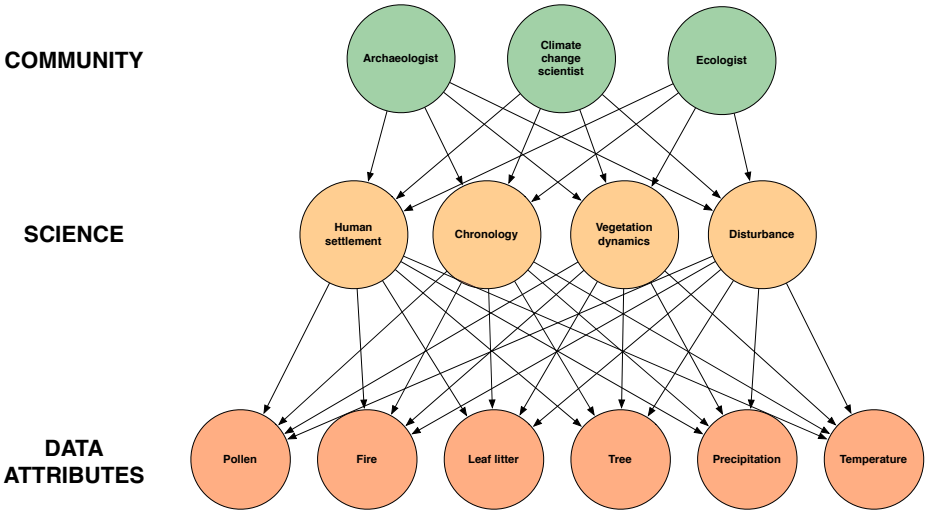


Fig. 7. Example of terms from DataONE metadata mapped to Bayesian network

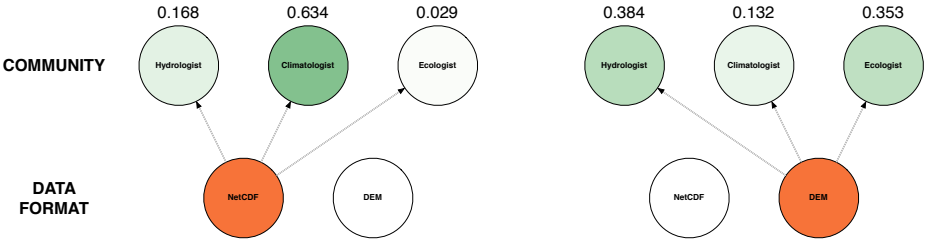


Fig. 8. Depending on the category of user different data formats will be more or less likely to be used in research

By finding the data with high relative likelihood, these probabilities can be used by data search applications to suggest potentially useful data for consumers who match community profiles, who are performing specific tasks, or working within specific scientific domains. Even with crude matching of terms to metadata text we begin to see value added in adopting this methodology. We anticipate being able to slowly build up richer descriptions of geospatial data, task and domain ontologies, and community space descriptions. Combined with Bayesian inference, we believe this holds great promise for new and better ways to find fit-for-purpose data.

4.2 Extending Toward Pragmatic Facets

Although the previous examples point to how we might use existing metadata to find potentially relevant and useful relationships between communities, tasks, scientific domains, and data schemata based on term matching; we contend that describing data

using relatively simple descriptive dimensions, such as those listed in Table 1, that target the pragmatics of data will provide additional valuable information for data discovery. Values along these dimensions can easily be assigned by data providers, consumers, and also third-parties, such as data custodians of spatial data infrastructures. Four description spaces consisting of two dimensions each and five ordinal values per dimension (0..4) form a universe of 390,625 possible descriptions, a tractable number for the Bayesian approach we advocate.

Conceivably, one could also develop alternative generative models that combine the data description based on pragmatic facets we propose with other commonly used descriptions of schema and file format. By measuring the utility of these various models with data *in situ* we can begin to evaluate and refine our data description methods in a systematic way.

5 Conclusion

Modern approaches to science are providing us with additional, non-traditional ways to describe our data, such as the way they are used, and the community they originate from. Currently, we cannot use these descriptions because they don't fit in our conceptual data models. Yet for us, describing data well is still a very complex, perhaps untenable — and certainly impractical — proposition. To take full advantage of these new descriptions, we need to let go of the need to define data universally and objectively. This is not how we *use* data.

A pragmatic approach to representation can allow us to preserve the value of current facts and ontological commitments (Peirce's firstness and secondness), but add in the notion of context where it is needed to account for the fact that many things are true only in certain situations or to certain groups. This paper provides a workable and flexible pragmatic model to describe data, which can be reconfigured according to need. We have demonstrated how some of the (usually) opaque knowledge about community, task, and domain can be inferred from current meta-data text descriptions — thus bootstrapping the movement towards richer descriptions without placing additional burdens of description on the data producer. We have a pressing need to evaluate the utility and practicality of all such new descriptions, along with the old, so we can know with some confidence where to focus our efforts when it comes to providing data descriptions. Our next paper will provide a practical assessment of utility and practicality by measuring improvement in search results when pragmatic aspects are facilitated in the search process.

References

1. Adams, B., Gahegan, M.: Emerging data challenges for next-generation spatial data infrastructure. In: Winter, S., Rizos, C. (eds.) Research@Locate 2014, Canberra, Australia, April 7-9, pp. 118–129 (2014), `http://ceur-ws.org`
2. Artz, D., Gil, Y.: A survey of trust in computer science and the semantic web. Web Semantics: Science, Services and Agents on the World Wide Web 5(2), 58–71 (2007)
3. Berry, B.J.: Approaches to regional analysis: a synthesis. Annals of the Association of American Geographers 54(1), 2–11 (1964)

4. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
5. Boisvert, E., Brodaric, B.: GroundWater Markup Language (GWML)-enabling groundwater data interoperability in spatial data infrastructures. Journal of Hydroinformatics 14(1), 93–107 (2012)
6. Carmel, D., Zwerdling, N., Guy, I., Ofek-Koifman, S., Har'el, N., Ronen, I., Uziel, E., Yogev, S., Chernov, S.: Personalized social search based on the user's social network. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, pp. 1227–1236. ACM, New York (2009)
7. Costello, M.J., Michener, W.K., Gahegan, M., Zhang, Z.Q., Bourne, P.E.: Biodiversity data should be published, cited, and peer reviewed. Trends in Ecology & Evolution 28(8), 454–461 (2013)
8. Crompvoets, J., Bregt, A., Rajabifard, A., Williamson, I.: Assessing the worldwide developments of national spatial data clearinghouses. International Journal of Geographical Information Science 18(7), 665–689 (2004)
9. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning 29(2-3), 103–130 (1997)
10. Egenhofer, M.: Toward the semantic geospatial web. In: GIS 2002: Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems, pp. 1–4. ACM, New York (2002)
11. Egenhofer, M.J., Frank, A.: Object-oriented modeling for GIS. Journal of the Urban and Regional Information Systems Association 4(2), 3–19 (1992)
12. Fegraus, E.H., Andelman, S., Jones, M.B., Schildhauer, M.: Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata language (EML) and principles for metadata creation. Bulletin of the Ecological Society of America 86(3), 158–168 (2005)
13. Gahegan, M., Agrawal, R., Jaiswal, A., Luo, J., Soon, K.H.: A platform for visualizing and experimenting with measures of semantic similarity in ontologies and concept maps. Transactions in GIS 12(6), 713–732 (2008)
14. Gahegan, M., Luo, J., Weaver, S.D., Pike, W., Banchuen, T.: Connecting GEON: Making sense of the myriad resources, researchers and concepts that comprise a geoscience cyberinfrastructure. Computers & Geosciences 35(4), 836–854 (2009)
15. Gahegan, M., Pike, W.: A situated knowledge representation of geographical information. Transactions in GIS 10(5), 727–749 (2006)
16. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening ontologies with DOLCE. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, pp. 166–181. Springer, Heidelberg (2002)
17. Goodchild, M.F.: Geographic data modeling. Computers and Geosciences 18(4), 401–408 (1992)
18. Goodchild, M.F.: Data models and data quality: problems and prospects. In: Environmental Modeling with GIS, pp. 94–104. Oxford University Press (1993)
19. Grira, J., Bédard, Y., Roche, S.: Spatial data uncertainty in the VGI world: Going from consumer to producer. Geomatica 64(1), 61–72 (2010)
20. Guarino, N.: Formal Ontology and Information Systems. In: Guarino, N. (ed.) International Conference on Formal Ontology in Information Systems (FOIS 1998), pp. 3–15. IOS Press, Trento (1998)
21. Heuvelink, G.B., Burrough, P.A., Stein, A.: Propagation of errors in spatial modelling with GIS. International Journal of Geographical Information System 3(4), 303–322 (1989)
22. Hobbie, J.E., Carpenter, S.R., Grimm, N.B., Gosz, J.R., Seastedt, T.R.: The US long term ecological research program. BioScience 53(1), 21–32 (2003)
23. Janowicz, K., Raubal, M., Kuhn, W.: The semantics of similarity in geographic information retrieval. Journal of Spatial Information Science (2), 29–57 (2011)

24. Janowicz, K., Schade, S., Bröring, A., Keßler, C., Maué, P., Stasch, C.: Semantic enablement for spatial data infrastructures. Transactions in GIS 14(2), 111–129 (2010)
25. Keim, D.A.: Designing pixel-oriented visualization techniques: Theory and applications. IEEE Transactions on Visualization and Computer Graphics 6(1), 59–78 (2000)
26. Kuhn, W.: Ontologies in support of activities in geographical space. International Journal of Geographical Information Science 15(7), 613–631 (2001)
27. Langran, G., Chrisman, N.R.: A framework for temporal geographic information. Cartographica: The International Journal for Geographic Information and Geovisualization 25(3), 1–14 (1988)
28. MacEachren, A.M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., Hetzler, E.: Visualizing geospatial information uncertainty: What we know and what we need to know. Cartography and Geographic Information Science 32(3), 139–160 (2005)
29. MacKay, D.J.C.: Information Theory, Inference, and Learning Algorithms, 7.2nd edn. Cambridge University Press, Cambridge (2003)
30. Michener, W., Vieglais, D., Vision, T.J., Kunze, J., Cruse, P., Janée, G.: Dataone: Data observation network for earth - preserving data and enabling innovation in the biological and environmental sciences. D-Lib Magazine 17(1/2) (2011)
31. Pearl, J.: Causality: Models, reasoning and inference. Cambridge University Press, Cambridge (2000)
32. Peirce, C.S.: The Collected Papers of Charles Sanders Peirce. Harvard University Press (1931)
33. Peuquet, D.J.: Representations of space and time. Guilford Press (2002)
34. Pike, W., Gahegan, M.: Beyond ontologies: Toward situated representations of scientific knowledge. International Journal of Human-Computer Studies 65(7), 674–688 (2007)
35. Raskin, R.G., Pan, M.J.: Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). Computers & Geosciences 31(9), 1119–1125 (2005)
36. Raubal, M.: Formalizing conceptual spaces. In: Varzi, A.C., Vieu, L. (eds.) Formal Ontology in Information Systems, Proceedings of the Third International Conference (FOIS 2004), pp. 153–164. IOS Press (2004)
37. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 3rd edn. Prentice Hall (2010)
38. Saalfeld, A.: Conflation automated map compilation. International Journal of Geographical Information System 2(3), 217–228 (1988)
39. Schwering, A.: Approaches to semantic similarity measurement for geo-spatial data: A survey. Transactions in GIS 12(1), 5–29 (2008)
40. Sen, M., Duffy, T.: GeoSciML: development of a generic geoscience markup language. Computers & Geosciences 31(9), 1095–1103 (2005)
41. Shi, W.: A generic statistical approach for modelling error of geometric features in GIS. International Journal of Geographical Information Science 12(2), 131–143 (1998)
42. Shvaiko, P., Euzenat, J.: Ontology matching: state of the art and future challenges. IEEE Transactions on Knowledge and Data Engineering 25(1), 158–176 (2013)
43. Sinton, D.: The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study. Harvard Papers on Geographic Information Systems 7, 1–17 (1978)
44. Sowa, J.F.: Syntax, semantics, and pragmatics of contexts. In: Ellis, G., Rich, W., Levinson, R., Sowa, J.F. (eds.) ICCS 1995. LNCS, vol. 954, pp. 1–15. Springer, Heidelberg (1995)
45. Worboys, M.F., Duckham, M.: GIS: a computing perspective. CRC Press (2004)