Investigation Design: The Structural Elements of Knowledge-seeking Efforts

Bryan Weaver[1] and Dieter Pfoser[2]
George Mason University
4400 University Drive, MS 6C3
Fairfax, VA 22030-4444
bweaver5@gmu.edu[1], dpfoser@gmu.edu[2]

ABSTRACT
Knowledge about human systems usually comes from deliberate, organized efforts. These efforts are increasingly collaborative with partnering among diverse teams of experts. This has spawned research in the means for data integration and data lineage, or provenance, to better enable sharing of data and workflows. Such research has focused on specific problems in domain representation, such as semantic domain modeling, provenance standards and methodologies, and automated workflow management. While all these contributions are highly relevant for knowledge seeking efforts, what is missing is a meta model that accounts for all elements of investigation. This work introduces the concept of investigation as a means to formalize knowledge seeking efforts involving collaborative human action. We model the investigation concept as a set of seven elements common to all knowledge-seeking efforts. Incorporated into the proposed investigation design are the concepts from emerging sensor-observation standards and W3C provenance standards. This design differs from other approaches for workflow modeling in its focus on reifying the management of effort – here referred to as a directive, and reifying analytic results – here referred to as judgments. Further, we provide and initial attempt to identify specific workflow and data model design patterns within each of the seven investigation elements. An example illustrates the various aspects of our approach.

Categories and Subject Descriptors
Design, Conceptual data models.

General Terms
Design, Knowledge Acquisition, Provenance.

Keywords
Linked Data, design, provenance, collaborative information systems, workflow, data management, investigation meta model.

1 INTRODUCTION

Understanding a continuously changing environment by collecting descriptive data is a challenging and resource intensive task. For one, the variety of modern sensor technology produces a wealth of "big" data whose sheer volume, velocity and variety (Laney, 2001) create a significant data integration challenge. Besides technical challenges, there are also organizational challenges. A complex data environment and constraints on an organization's resources often necessitates collaboration beyond traditional organizational boundaries. Metadata describing the data collection and analysis effort is needed to maximize the use and reuse of the collected data across often-complex social networks. So what is missing? Team building processes and the definition of respective procedures and organizational standards have characterized collaborative efforts. Efforts centered on User-Generated Content (UGC) and open data popularized self-organization as an efficient approach to collaborate efforts. Consider here for example Wikipedia and its map data peer effort OpenStreetMap. What is common in both approaches is that they have a very narrow focus and stringent task descriptions supported by specific tools to further reduce the complexity of tasks, e.g., editing text documents based on the user's knowledge (Wikipedia) and digitizing satellite imagery to compile and enrich map datasets (OpenStreetMap).

User-generated content around specific goals and objectives provides us with great examples that shared understanding facilitates successful teamwork and collaboration. In non-UGC contexts, efforts to facilitate collaboration have been limited to a shared understanding of the data. The key concepts that come to mind are ontologies (Staab & Studer, 2004) to denote the meaning and structure of concepts and, more recently, Linked Data (Heath & Bizer, 2011), which represents a method of publishing structured data so that it can be interlinked and become more useful to a greater user group. It builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages for human readers, it extends them to share information in a way that readable by computers. This enables data from different sources to be connected and queried using standard WWW technology. To move beyond the sharing of datasets to collective development and sharing of knowledge, one needs to formalize and document the aspects of the knowledge acquisition effort that can be made explicit, i.e., its purpose, structure, and findings. Here, we use the term *investigation* to refer to such efforts. We define *investigation* as an intentional effort, of any scale or scope, to acquire knowledge.

Trusting results and understanding the appropriate use of information created by social science investigations remains challenging. Too few research results are published with links to supporting workflows and data (Davidson & Freire, 2008). When datasets are made available, they vary greatly in the models used to convey the relevant information others need to make proper use of the content. Further, most available datasets do not contain sufficient context of the perspective, boundaries, or rules of the study from which its data were born – making it difficult for others to evaluate its fitness for use.

In spite of these hurdles, the *demand to share data and integrate workflows across research teams is growing* rapidly (National Academies, 2004). The social complexity of investigations and the need for good data demands a re-examination of how the discovery process for social science is led and organized. In addition, research is becoming increasingly interdisciplinary. There are four drivers to this: (i) the complexity of nature and society, (ii) the desire to explore problems not confined to a single discipline, (iii) the need to solve societal problems, (iv) and the power of new technologies (van Rijnsoever & Hessels, 2011). For example, The Allen Brain project uses multi-disciplinary teams to discover and map the functions of the human brain, for example (Kandel, Markram, Matthews, Yuste, & Koch, 2013). To be successful, such investigations require good design and coordination to overcome the individual motivations and cognitive distance between parties (Nooteboom, 2000). The complex social nature of investigations now extends beyond professional communities. The use of open tools and social media has enabled crowd-sourced data to contribute meaningfully to well-designed research (Silvertown, 2009).

The *demand for variety and volume of data inputs to address social problems is increasing* (National Academies, 2004). This requires greater cooperation within disciplines to share data and methods. A 2015 Open Geospatial Consortium (OGC) white paper argues "domain specific but technically interrelated IT standards for communication and data integration within and between domains" is necessary to address the many environmental challenges facing society (McKee, 2015). The social complexity of research demands a means for managing goals, roles, and constraints within investigations. The breadth of data collection needed to address these large problems requires workflow design conventions and data standards, which enable collaboration at a grand scale.

Several frameworks have been proposed for maintaining explicit knowledge of complex domains (Janowicz, 2012; Shaon et al., 2012), for knowledge engineering methodologies (Studer, Benjamins, & Fensel, 1998), and for data lineage preservation, which is necessary to integrate workflow across teams (Rajendra Bose & Frew, 2005; Car, 2013a; Freire et al., 2006a; Freire, Koop, Santos, & Silva, 2008a; Li, Dragicevic, Veenendaal, & Brovelli, 2013). In addition to the above frameworks for creating and maintaining domain models, a number of standards were created in the past few years to better establish interoperability of geospatial data. The Sensor Web Enablement (SWE) suite of standards adopted by the OGC describe models for recording and serving sensor data and observations (Botts, Percivall, Reed, & Davidson, 2006; Bröring et al., 2011). To date, however, there has been no complete data framework or set of standards for representing all aspects of an investigation. Particularly absent is the reification of the investigation itself.

In short, considering the complexities of social problems, the institutions addressing them, and the big data era, we must re-examine the basic workflow and data modeling patterns that enable collaboration across organizations and their respective data environments. The proposed investigation design incorporates leadership functions with existing frameworks and standards to present a meta model for the structure of an investigation. We argue that all investigations have a set of common elements, which form its structure. An e*lement* is defined here as a general category of data necessary for carrying out an investigation.

This paper has four remaining sections. Section 2 provides the motivation for creating an investigation meta model and evaluation of existing approaches. Section 3 defines the investigation meta model by detailing its structure, while Section 4 provides a fictitious example. Investigations, which thoughtfully reflect the proposed elements, can provide the information necessary for conducting complex, integrated efforts while providing the

context necessary for content reuse in unplanned ways. Conclusions and future work are presented in Section 5. A running, fictitious investigation scenario illustrates the contrasts between existing works and the proposed investigation design.

2 MOTIVATION AND RELATED WORK

A motivating example highlights the current challenges of complex, multi-party, social science investigations. Addressing these challenges, we argue, requires the application of three key principles reflected in, both, the presented investigation structure and the organizational execution of such projects.

The example used throughout this paper uses a fictitious wildlife conservation scenario. Assume that a national effort has two goals for increasing its elephant population: 1) reducing poaching, and 2) developing a habitat conservation plan. The National Park Service is charged with overseeing and delegating responsibilities for achieving these goals among its many other investigations.

2.1 Motivation

Good decision-making begins with understanding the goals of one's efforts. Understanding the goals makes it easier to adjust procedures or data sources when unforeseen hurdles or new opportunities arise.

To be a partner in a knowledge-seeking endeavor requires knowing not only one's role and workflow, but also knowing the roles of others. This is particularly true when there exists a workflow dependence. Partners must be able to know their dependents and dependencies across the investigation. For example, anti-poaching patrols within a park might rely on the real-time observations created by biologist from their network of acoustic sensors. Likewise, the biologists might depend upon the field observations of anti-poaching teams to help them make sense of elephant response to threats. If the acoustic sensor system stops working, the anti-poaching patrol needs to be aware so they can adjust their workflow appropriately. If the poaching patrol replaces place name spatial attribution with Global Positioning System (GPS) coordinates, other teams in the investigation might need to adjust their workflows to account for poaching patrol observations with coordinates.

One also must understand the constraints placed upon them regarding standards, methods, or domain-specific conceptualizations. The biologist team and the anti-poaching team likely need to refer to the same controlled vocabulary in describing poaching threats and events. They need to have a common understanding of the duration of the study and the spatial extent of the study. Does the investigation cover all national parks? Does it extend to private land? Do I report when there are no threats observed or only when I see a threat? As investigations proceed, goals and priorities can change for various reasons. Therefore, teams must be able to accept changes to roles, rules, and goals by leaders of the investigation or parent investigations. Finally, managing interdependent workflows across, potentially multiple investigations can be challenging. Teams need to optimize workflows to meet all obligations across all investigations.

All of these challenges illustrate why socially complex investigations are difficult to execute. How can multi-party investigations agree and work towards common goals despite differences in motivations? How can subordinate investigations adhere to the demands of higher-level investigations and interdependencies with peer investigations? The answer to these questions is both organizational and structural. Underpinning the proposed design are the complimentary organizational principles of leadership and empowerment and the data lineage principle of provenance.

- Collaboration requires good *leadership* to guide and integrate the efforts of diverse organizations toward its common goals.
- Workflows are most effective and adaptable when those charged with executing them are *empowered* to compose them.
- *Provenance* is the enabling data principle to achieve trust, accountability, and efficient integration of workflows.

The remainder of this section highlights published conceptual models or standards for representing the data and process of collaborative knowledge-seeking investigations.

2.2 Related Work

Several frameworks and standards exist for recording data pertaining to a problem domain. There is less work discussing an overall process for generating data in collaborative effort. A review of published works is presented, beginning with the more concrete aspects of domain-related data and concluding with the more abstract problem of representing the effort of investigating.

2.2.1 Sensor Data and Observations

Data about a domain problem is obtained through data collection activities. The most basic data about a domain is obtained through direct observation of the subject. The OGC and ISO Observations and Measurements standard (O&M) and the Semantic Sensor Network ontology (SSN) are two established conceptual models for representing data that measures the environment and relates those measurements to objects within a domain. Sensor data is the most objective measure of the environment. SSN uses the term *stimulus* to refer to the data of a measurement prior to interpreting the measurement to a specific application domain (Compton et al., 2012). O&M does not explicitly model *stimulus* data distinct from observation data but relates an observation to a process which can represent a method or sensor used to create the observation (Compton et al., 2012; Cox, 2013). In both models, sensors can be machines or human witnesses.

The O&M standard is a conceptual schema for storing observations derived from sensors. According to this standard, an observation, is "an act of observing or otherwise determining the value of a property (Cox, 2013). Figure 1 illustrates the conceptual observation model in the OM standard.

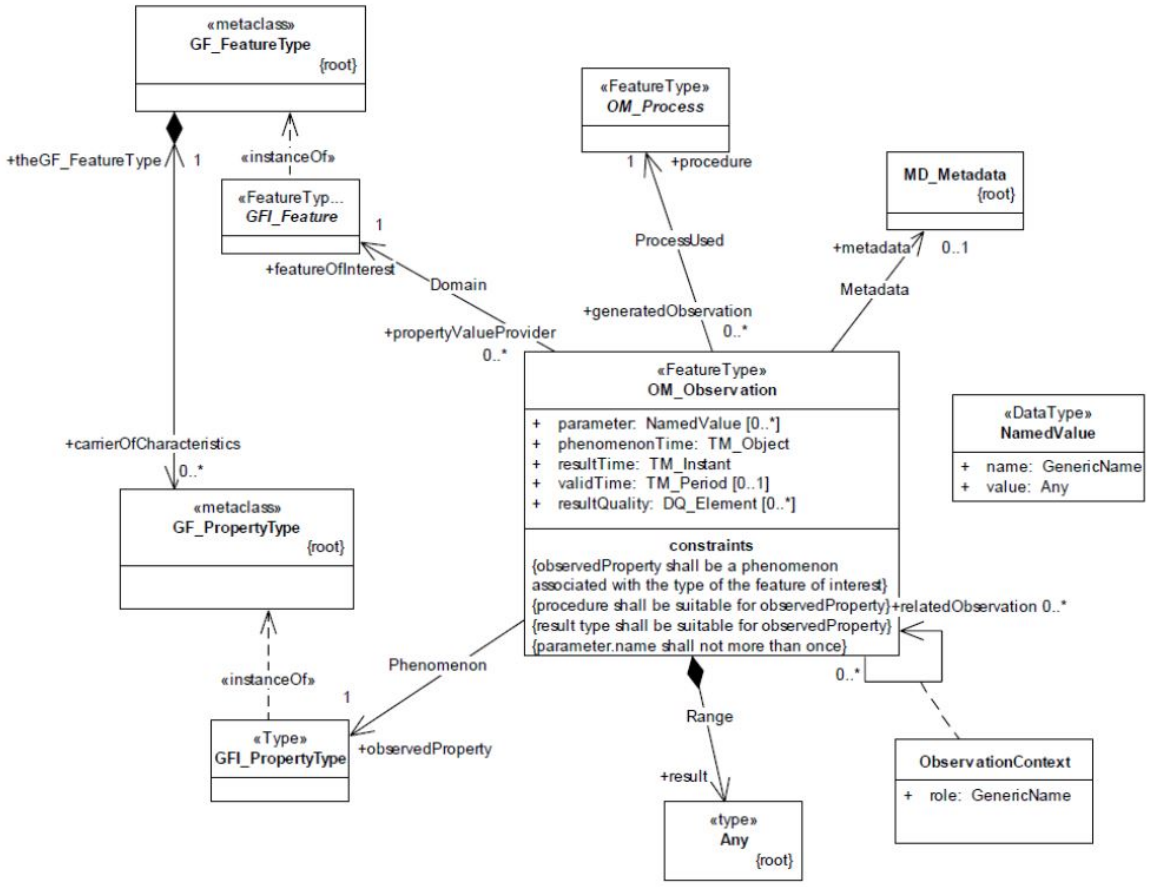

Figure 1. ISO 19156 Observations and Measurements Observation Type (Cox, 2013)

The O&M schema defines several classes for different facets of an observation. An *O&M_Observation* references a feature (object) from a domain model and through a process assigns a value to a property for the referenced object. The *OM_Process* relates the sensor equipment to the observation (Cox, 2013). The Climate Research Unit within the University of East Anglia describes their use of this model for integrating a variety of geospatial observation data (Shaon et al., 2012).

SSN provides a similar sensor-observation modeling pattern as the O&M schema. Central to the SSN ontology is the Stimulus-Sensor-Observation pattern, depicted in Figure 2. In the SSN ontology, sensors are defined as "anything that senses", including hardware devices, people, or systems. Sensors detect stimuli. Stimuli, in this context, are the things in the environment the sensor can measure (Compton et al., 2012). Observations result from

using a sensing method to identify an instance of a feature property from stimuli. Several approaches have used the O&M or the SSN models to better standardize observation data across large sensor networks and investigations.

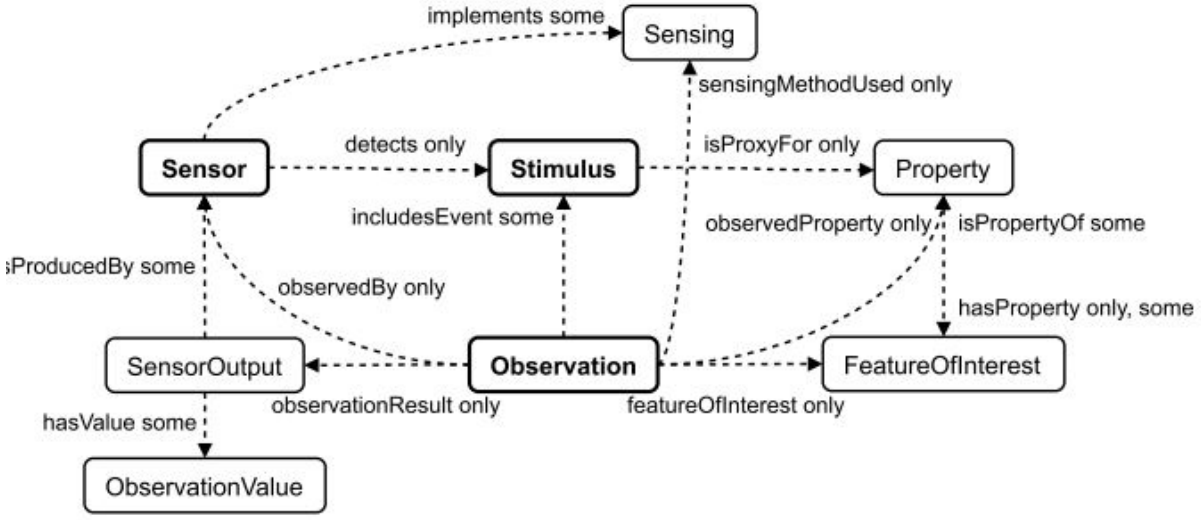

Figure 2. The Stimulus-Sensor-Observation pattern on the SSN ontology (Compton et al., 2012)

2.2.2 Referencing common features: ontologies

Observations are usually a means to an end. They are the empirical fuel in the analysis of the problem domain. A problem domain is represented as a set of objects, or features, their properties, and relationships to other objects. Increasingly, domain models are formalized as ontologies. Data models, unlike ontologies, are designed for a specific application, thereby constraining its reuse. Ontologies, on the other hand, allow for domain concepts to be defined and constrained but also used by several applications (Spyns, Meersman, & Jarrar, 2002). This is important for interorganizational efforts, where specific implementations and uses might vary. However, creating (and maintaining) ontologies shared across organizations is difficult. It requires a complex process to negotiate ontology alignment while maintaining individual organizational interests. The DOGMA-MESS is a framework for interorganizational ontology engineering based on continual, increasingly complex, common ontology alignment pattern (de Moor, De Leenheer, & Meersman, 2006). By contrast, Janowicz argued for an approach to ontology engineering based on a grass-roots observations he called Observation-Driven Geo-Ontology Engineering (ODOE). His approach argues that provenance data of observations applied across participants can be used to semi-automatically derive a top-level ontology that is common across all participants without violating the assumptions of each participant (Janowicz, 2012). Regardless of the approach to build an ontology, collaborative efforts require common conceptions of the domain across participants.

2.2.3 Provenance

The principle of data provenance pervades recent designs for observation data, domain models, and the processes to create data. The term provenance is defined as the origin or source of something (Lakshmanan, Curbera, Freire, & Sheth, 2011). Users of information integrated from various sources need to understand its provenance in order to trust it. Provenance data is recorded at the feature level – fine grained, or at the dataset level – course grained (Harth & Gil, 2014).

Traditionally, provenance metadata is more often provided at the dataset level, describing aggregations of observations. For example, ISO 19115:2009 is the Geographic Information Metadata standard. This standard defines the schema required for describing geographic information and services, including information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data. One problem with the dataset metadata approach is that the information is only retrospective – created only a posteriori – and does not address workflow management data which directs the behavior of data producers. Secondly, it is cumbersome to maintain. Citing its complication and the general neglect of ISO 19115:2009 in practice, a recent OGC study decided to recommend the PROV model for recording data lineage for geospatial data (OGC, 2014).

Rigorous provenance standards for workflow activities enable their precise replication. Research on workflow replication has been a recent research focus (Rasjendra Bose & Frew, 2005; Car, 2013b; Freire et al., 2006b; Freire, Koop, Santos, & Silva, 2008b). Prospective provenance of activities captures the sequence of tasks that must be followed to produce a specific outcome. Retrospective provenance captures the steps executed as well as information about the processing environment used to create a specific product (Freire et al., 2008b). There are a few products that capture prospective and retrospective provenance. VisTrails is a system that captures provenance information for both workflows and data products (Freire et al., 2006b).

2.2.4 Toward Investigation Design and Structure

Integrating workflows across multiple parties toward common domain knowledge goals requires more than observation data and domain data design, it requires management of the collaboration process. Management activities include planning and guiding the execution of investigation activities. The outcomes of these activities are plans, or directives, for participant action. Plans are themselves data and need to be reified, as noted in PROV (Missier, Belhajjame, & Cheney, 2013). Once reified, plans are referable as provenance metadata. Retrospectively, the plan itself provides important context to users of the resulting data. Prospectively, a plan guides actions of the participants. Works from the knowledge based systems and project management fields partially address how plans can be modelled for investigations.

Design methodologies for knowledge based systems help form the creation of a general investigation design. CommonKADS is a popular methodology for knowledge engineers to design knowledge base systems. The CommonKADS method facilitates creation of a knowledge system design through documentation of Context Models (Organization, Task, Agent), and Concept Models (Knowledge, Communication) (Schreiber, 2000). While employing a CommonKADS methodology can aid in the design of a specific knowledge system implementation, its output is not recursive. Further, it does not provide an overall pattern or structure for investigating, such as observation data design. Pomponio and LeGoc point out that knowledge engineering (KE) models, such as CommonKADS, do not address the data mining processes of model development, generally referred to as knowledge discovery in databases (KDD). They suggest a framework to reduce the distance between and experts problem solving model and KDD models using Timed Observation Theory (Pomponio & Le Goc, 2014).

Policy management and project management research offers themes of data, which are necessary to represent in a plan. After all, investigations can be considered a knowledge project undertaken within specific policy conditions. The Project Management Body of Knowledge (PMBOK) provides the core concepts of project management (Project Management Institute, 2008). The PROMONT ontology, for example, formalizes the established project management concepts in the German project management norm DIN 69901 data model (Abels, Ahlemann, Hahn, Hausmann, & Strickmann, 2006). Policy conditions include the expression and enforcement of obligations. Obligations are defined as actions that users are required to take or states of affairs which must be maintained (Elrakaiby, Cuppens, & Cuppens-Boulahia, 2012). Formal definitions of classes, properties, and relationships of project and policy management concepts helps enable integration and communication of investigation functions across distributed teams.

Despite robust models for observation design, domain ontology representation and engineering, and methodologies for knowledge system design, no existing model provides a comprehensive approach to structuring all necessary elements of contemporary investigations, described above as mutable, recursive, and collaborative. Of particular importance is the lack of formalization for directing the overall process of generating data. The elements of an investigation include observation models, domain models, and the reification of the means to populate such models. The remainder of this paper details our proposed structural design for collaborative investigations.

3 INVESTIGATION STRUCTURE

Investigation elements concern the representation of knowledge about a domain. This design presents seven elements common to all knowledge-seeking efforts. Supporting this structure are a set of complimentary principles. Figure 3 illustrates these principles.

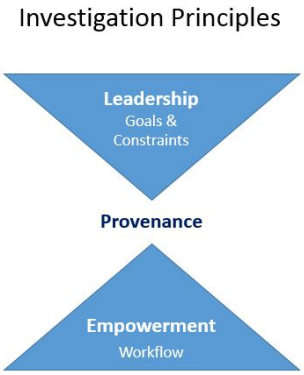

Figure 3. The principles of leadership, empowerment, and provenance underpin investigation design.

Investigations require leadership to align contributor efforts toward common goals. Leaders must also empower contributors to create and execute workflows, which meet the demands contributors face from multiple investigations. For example, leadership from the National Park Service might encourage and align efforts between government, non-profit conservation organizations, and the public. National Park Service leadership might constrain these contributors to providing observations that meet a specific data standard. With the exception of these constraints, contributors ought to be empowered to create workflows that meet the needs of all investigations for which they are participants. Provenance is the data lineage principle, which links leadership direction to the direction from higher level, parent investigations to the data created by contributors. Provenance also enables the presentation of evidence chains supporting the results of an investigation.

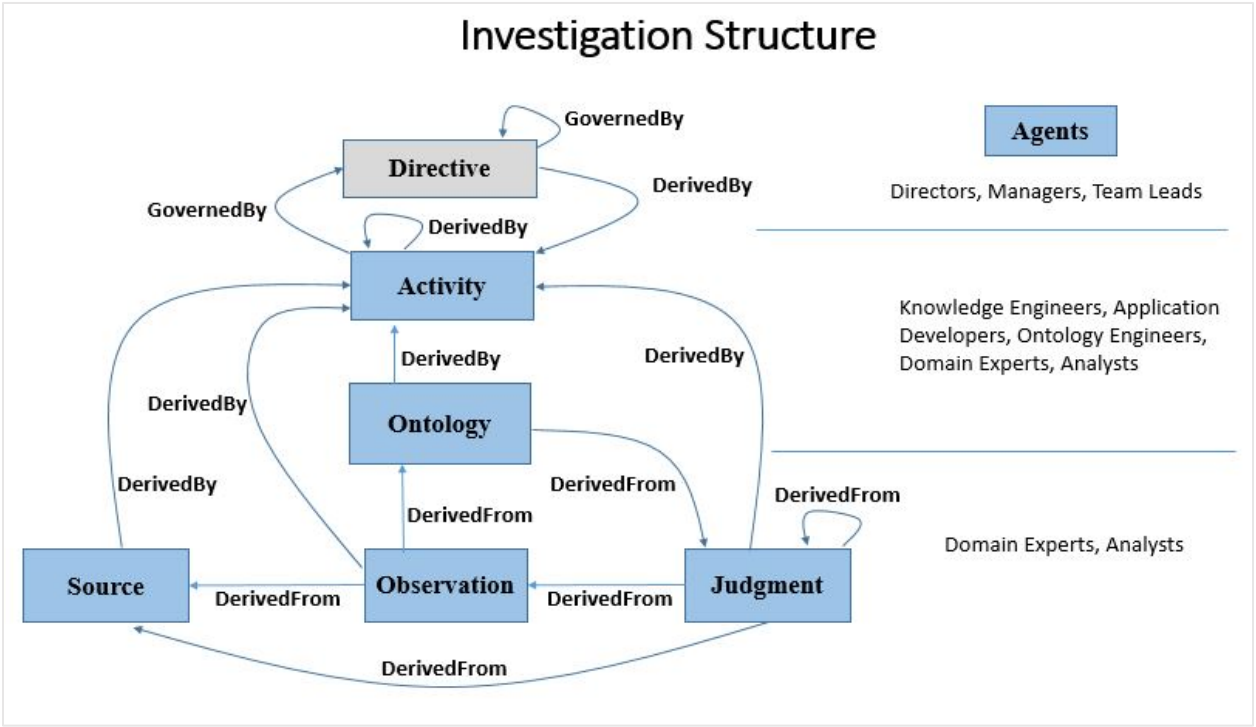

Figure 4: Investigation elements and their relations. Agent positions (right) generally correspond to investigation elements (left) from leadership (top) to domain analysis (bottom).

3.1 Investigation Element Relations

The basic relationship among content elements is shown in Figure 4. An investigation is an intentional effort, of any scale or scope, to acquire knowledge. Investigations are comprised of seven elements. A *directive* is the element within an investigation, which provides leadership information and declarations which govern the activities of an investigation. A *directive* is defined as a formalized description of the goals and constraints of an investigation. An *activity* is a process, or action, used to create content. No data is created without an activity. The *ontology* refers to the knowledge models used within an investigation. It is an abstraction of the real world processes pertinent to the investigation and the metadata concepts used to describe investigation data. Ontologies are maintained by activities which generate semantic judgments, or changes to the working model of the problem domain. At the bottom left of Figure 4 is the *source* element. A *source* is a uniquely identifiable artifact representing a direct measurement of the environment or subject. *Observations* build off sources and the ontologies used. An *observation* is the observed instance of a class, subclass, or property within an *ontology* derived through the application of an *activity* on a specific set of *source* data. *Judgment* is an assertion resulting from the application of an analytic *activity* to all available and relevant data.

People and activities pervade all elements of an investigation. People carry out or are accountable for every activity and no data is created without an activity. People, and organizations of people, are referred to as *agents*. The complete set of *agent*s participating in an investigation are often referred to as a community of interest (COI). Provenance metadata is applied directly to the individual instances of each of the seven content classes. Therefore, in this model, data is attributed to its *DerivedBy* and *DirevedFrom* inputs. There is no need for a conventional metadata document summarizing the data contained within an investigation. For example, to list all imagery sources used to create observations of elephants, a user should simply query for all unique image sources linked to observations of the elephant feature class.

It is important to consider the recursive nature of this investigation design. An investigation might evolve into several, progressively more tactical, investigations as complexity is revealed in the problem domain. For example, the National Park Service might soon realize that the interaction with private landowners with wildlife is a key aspect of conservation planning. This might spur another investigation with the goal of understanding the interaction of ranchers with wildlife. Conversely, investigations can be nested within multiple, higher-level investigations. For example, a team of conservation biologists might be part of an investigation of both elephant social patterns and the elephant conservation investigation. They would likely manage their workflow so that each observation captures the data necessary to support both investigations. The remainder of Section 3 defines each content element in more detail.

3.2 Investigation Element Description

Each element is described in this section. The concepts are described first with a definition, followed by a statement of purpose, justifying why the content concept is a necessarily distinct element. Design considerations of each element and the key relationships among elements are also discussed. Finally, paths for implementing each element in the creation of an investigation workflow is briefly addressed.

3.2.1 Agents

Agents are the people and organizations of people involved in an investigation. There would be no investigation without at least one *agent* responsible for the effort. Every activity is linked to an *agent*.

Purpose – it is important for investigations to maintain the data links between the content created and the creator of the content. Ultimately, people are responsible for the activities and outcome of an investigation. Also, the activities used to create information often require human thought processes which are not easily reified. A unique identifier representing the *agent* allows for relating the 'thinker' to the information outcome of human reasoning, thus maintaining some sense of data provenance.

Design Considerations – representing people within the data of an investigation allows for *agent* constraints to be implemented in the content creation process. In large endeavors, there are possibly several different *agent* roles. Each role might have unique constraints on the content the role can access and author. This is the primary design consideration for *agent*s.

Individual Roles – customers, senior managers, knowledge engineers, project leads, and analysts play different roles in an investigation. The specific typology of roles is determined by the organization which oversees the investigation. Customers, or stakeholders, often might have specific roles that enable query of a content and creation of specific request data, but prevent them from creating domain observations and judgments. Managers might be primarily responsible for defining the goals and rules for conducting an investigation. Knowledge engineers might have the unique ability to change a metadata ontology for the investigation. Project leads might have specific quality

control privileges and analysts might have the lion's share of the work collecting sources, making observations, and analyzing observations to create judgments. Considering what data the roles within an organization are meant to create, or not create, should be reflected in the attributes of each *agent*.

Organization Roles – much like the qualities of individuals, people grouped for a common purpose, an organization, might have attributes that are inherited by all members of the group. For example, a system might constrain any politician from creating national security or intelligence observations or judgments in order to prevent the politicization of intelligence content. The system might also allow any politician to draft a goal or request for intelligence. Similarly, a system might constrain analysts of a specific discipline from changing the domain ontologies for which they are not a certified member of the appropriate community of experts. Park police are likely not qualified to change the ontology that models elephant biology. That privilege might be reserved for individuals specified by Chief Wildlife Biologist.

Dependent Elements – for *agent*s to exist, they are not dependent on any other element. However, for *agent*s to create and execute an investigation, *agent* and activity are co-dependent. In other words, an investigation cannot be defined or executed without activities, and an activity cannot be executed without a responsible *agent*.

Agent
what it is: people and organizations *what it does:* invokes or performs activities to create data
key ideas: Managers orchestrate Technologists provide tools Experts execute

Figure 5: *Agent* element summary

3.2.2 Directive

A *directive* is a formal description of a knowledge-seeking effort. Every investigation is represented by one and only one directive. Directive data provides information which defines the boundaries and rules for participation in the investigation. Through formal semantics, the specific domain ontologies, the activities or methods required, and the rules for collecting and storing data can be defined. A directive can be thought of as a combination of a "plan" from the PROV ontology and metadata describing the context of a project. Since investigations are often nested within other investigations, therefore, directives can be nested within other directives. This allows for a hierarchy of investigations, often necessary in complex research, with each successive lower level investigation providing more local, specific direction to smaller, more homogeneous aspects of a study.

Purpose – directives serve to reify the investigation. It is important to capture this content for two reasons. First, providing clear direction to participating teams through data improves how teams work together to integrate their respective outputs toward the stated goals. Second, the context provided by a directive can be essential in determining the fitness of use of data resulting from the effort toward other purposes.

Design Considerations – The information in a *directive* defines the purpose and constraints of the investigation. To enable modular authorship of rules pertaining to specific elements of an investigation. Table 1 shows an early attempt at the properties associated with the *directive.*

Goals and Purpose – the activities of an investigation are guided foremost by the knowledge-seeking goals of the investigation. Directives should convey what specific problem shall be addressed by subsequent investigation activities. Hypothesis, goals, and objectives should be formalized. Result requirements should be formalized as well. The Elephant Conservation Plan directive, for example, might stipulate the expected outcome is a map showing recommended expansion of elephant habitat and the associated costs.

Key Assumptions – a common understanding of the most important assumptions can be critical for comparing results across investigations. Explicitly stating key assumptions also allows participants to challenge them if presented with contradictory evidence. For example, a key assumption of the Elephant Conservation Plan is that more protected area will increase the size of the elephant population. The Chief of Park Police might argue, however, that more dispersed elephant population makes them more difficult to protect with the same number of anti-poaching police. Stating key assumption provides a common starting point for research, and one that can be challenged by participants.

Management – adjusting to changes in investigation conditions often requires management actions. We have addressed that directives can be used to communicate changes in the conditions of an investigations. They can also be used to communicate leadership responses to change, such as altering objectives, the domain ontology, personnel

roles, or processes. Delegating responsibilities is another management function. A directive might assign a specific agent in charge of managing and mandating the set of data standards for the investigation, for example. Finally, management also makes decisions for measuring success. Directives can be used to formalize success criteria. Considering what management functions leaders of the investigation might need to formalize is beneficial.

Essential standards – to enable interoperability of content, standards often need mandating across the investigation. Standards can include data models, ontologies, and metadata. They can also include procedures for analyzing data or communicating across organizations. A "Be on the lookout" expression could be standardized to serve as an alert mechanism to all participants. For example, if a specific set of identifiable poaching vehicles are discovered, the National Park Service could disseminate a "Be on the lookout for" instruction to all subordinate investigations. Ideally, the directive would communicate an unambiguous signature for the vehicle objects and specify a search priority for subordinate investigations.

Boundaries – all investigations are bounded by a time, geography, and the set of objects within a problem domain. Directives can be used to establish common boundaries across all teams involved in the investigation.

Collaboration Strategies – investigations are often conducted within formal organizations with administrative control. In these environments, a directive serves as a requirement to create knowledge. Increasingly, investigations involve partnerships among two or more formal organizations and sets of independent agents. In these environments there can be less clear lines of accountability for contributors. Formal agreements for investigations, such as memorandum of understanding or joint ventures, can identify the roles and responsibilities across participants. Generally parties must have mutual interests and synergies to overcome the cost of alignment or be directed by an overarching authority that assures mutual commitment (empirical collective action theory).

Dependent Elements – because a directive can be the initial content commencing an investigation, it is not dependent upon any other element besides an activity that creates the directive.

Directive
what it is: reification of an investigation *what it does*: means of directing activities to achieve goals
key ideas: formalize goals and activity constraints conveys context of effort to users of data recursive and mutable

Figure 6: Directive element summary

3.2.3 Ontology

An ontology is an "explicit specification of a conceptualization". A conceptualization is "the objects, concepts and other entities that are presumed to exist in some area of interest and the relationships that hold among them (Gruber, 1995). It is important for any investigation to first have a model representing what is known about the domain prior to collecting data to refine what is known. Representing the problem domain model should be complimented by a common metadata model for describing information about data records. This investigation design can be considered a meta model for this second type of ontology. To achieve provenance to original evidence sets, it is sometimes necessary to represent the physical objects that enable the existence of domain information. For example, the field notebooks used by field biologists need to be represented as data objects in order to maintain lineage of observations which cite a field notebook as a source. Further, it is important that these three types of models be represented such that a computer systems can make inferences about the model itself as well as the instances that comprise the related real-world observations and judgments. Therefore, the concept of an ontology, a semantic model, is used here to represent the domain model referenced by an effort.

Purpose – the ontology element serves as a common model for the participants in an effort to communicate and maintain data about a domain. It provides a controlled vocabulary for discourse concerning all participants in the effort. This is important in reducing misinterpretation of observations and judgments in the problem domain, expressing ideas, and creating seamless workflows across teams.

Design Considerations – as introduced above, there are three types of ontologies that should be consider in creating an investigation.

Domain ontologies – Domain Ontologies are semantic models for describing the subject of investigation. It is important to have a sufficiently precise model of the real-world system to which the observations and judgments of the investigation will be referenced. Signatures, or distinguishing characteristics of an object property, object

instance, or object class, are properties of objects in an ontology. Signatures can be explicit or implicit. Explicit signatures are formally defined, documenting a specific pattern of primary source data traits as sufficient evidence for identifying the property, object, or object class. Implicit signatures are best described as primitives. These signatures are not formally defined. If an object or object property has an implicit signature, the *agent*s within the investigation are trusted to record observations of such objects unburdened by specific source data attributes.

*Metadata Ontologies – M*etadata Ontologies are semantic models which define how data elements are described with data. For example, a metadata ontology would define the provenance attribution that is ascribed to the instances created in the domain ontology.

Physical Dependence Ontologies – physical Dependence Ontologies are the semantic models that describe physical and virtual resources used in the investigation. The configuration and content representation of computer networks and physical data storage devices ought to be considered in an investigation. For example, a physical dependence ontology could define how the physical notebook (made of paper and ink) containing an interview with a research subject is related to the digital source of the interview content.

Ontology Composite Set – it is important for investigations spanning many teams and domains to identify a set of common ontological elements to which all teams will refer. This often will cross multiple domain ontologies. This is necessary, particularly in the social sciences, because many investigations require the participation of teams working in different domains, or with competing perspectives, but with some substantive overlap. A team of remote sensing specialists mapping potential elephant conservation environments might require an *ontology composite set* that references classes in uses both a zoological ontology and a landuse ontology.

Dependent Elements – the Ontology element is dependent upon Judgment, Activity, and Directive content elements. Judgments asserted by domain experts form the ontology. These assertions can be born from inductive or deductive reasoning and an activity, or method, for changing the ontology. Put another way, a community of domain experts assert the necessary existence and inclusion of an entity and its properties through a governed application of top down logical deduction or through the examination of data, which leads to an evidence-based assertion. Ontologies, thereby, form the model for representing what is thought to be true and relevant about a domain. They are modified when an ontological judgment is made through a semantic activity in an investigation. More extensive discussion on ontology maintenance is beyond the scope of this paper. That said, an investigation would likely define the rules and methods for maintaining the relevant ontologies.

Ontology
what it is: explicit description of problem domain *what it does*: enables unambiguous reference to concepts
key ideas: domain, metadata, and physical dependence types instances created through observation and judgment created by semantic judgments

Figure 7: Ontology element summary

3.2.4 Activity

An activity is something that happens over a period of time and acts upon content. For some data to be generated, there must be an activity and input. Activity can be data representing physical process, such as a meeting of people to document an ontology. More often, an activity refers to an algorithm that creates more refined data from less refined data. For example, a software algorithm may create an observation an elephant within 1 kilometer of an in situ acoustic sensor, given the input of a specific frequency and an acoustic signature library of elephant vocalizations.

Purpose – it is important to relate the processes used to create data to the data itself. This makes the results of a process much more useable. Most generally, activity lineage helps users understand the data of an investigation. The reproducibility of an investigation is not possible without identification of all key activities. Also, clear process lineage enables other investigations to determine if the processes or data can be repurposed for their use. Activity lineage also enables identification problems and reprocessing of content that was generated by a faulty process. Relating data to the activities from which they were born is a fundamental principle of data provenance.

Design Considerations – activities can be examined and defined in two ways: by the instruments that perform the activities, and 2) by the functions they perform.

At the highest level, activities can be considered human-based or machine-based. *Human-based activities* are those activities performed by humans, whose cognitive processes are often irreducible into data components. Examples of human-based activities include planning, meetings, discussion, evaluation, or personal observation. It is difficult to model the substance of a meeting or discussion, beyond the purpose, scope, and decisions. The paths of reasoning and the intellectual methods applied to argument or human interpretation often must be considered a "black box". Often only the input data sources, reasoning agent (human) and output data – transformed from the "black box" mental process – are cost effective to represent in data.

Machine-based activities are procedures executed by computers or computer systems. Examples of a machine-based activities include sensor software, automated object detection, and speech-to-text software. Of course all machine-based activities have a human author and an actor responsible for executing the procedure. Unlike human-based activities, machine-based activities are typically reproducible – given identical access to inputs - because they are based solely on logic. Therefore, complete provenance to machine-based activities is achievable if all variables of a machine execution are recorded. This includes hardware, software, and network parameters. For this reason, it is recommended that software development conventions include capabilities for creating audit trails for all variables values of an execution. Such a machine-based activity execution audit enables the reproducibility of a machined based process, assuming the data, hardware, and network resources are available. There are trade-offs between human and machine activities. Humans are more adaptable, but machines are faster and more consistent at logic. *Composite human-machine activities* take advantage of the strengths of both machines and humans. An example is a software program that has a user interface for a human to input variable parameters. The human can adjust the parameters based on the immediate context, and a computer can calculate the results of the input parameters efficiently, without error. The inherent, ephemeral qualities of human thought versus the formulaic qualities of their inventions must be considered in the design and implementation of activities.

Functional typology of activities – activities can be classified by the functions they perform. Every element of this investigation design is implemented with a set of activities. At the highest level in a functional typology of activities is an *Organizing Activity*, defined as an activity for managing activities. At the outset of any endeavor, people determine how they are going to organize themselves and use their resources to accomplish their goals. Organizing activities constitute part of the provenance for an investigation. In other words, there is some activity that creates an endeavor formalized as an *investigation*. Various types of activities created domain-related data. A suggested typology of such activities follows.

Semantic Activities – semantic Activities are processes used to create and modify the ontologies used within an investigation. An example of a human-based semantic activity is a meeting of a domain ontology matching team, where the meeting is the activity with specific goal, scope, and outcome properties. An example of machine-based semantic activity might be the execution of a software application nominating a new subclass of butterfly species based on a divergent pattern of observations within a species of butterfly.

Signature Activities. Part of developing and applying an ontology to a real world data is the development of rules for instantiating what is observed to the semantic model. A signature in this context refers to a distinguishing set of characteristics which sufficiently distinguish an object or a set of objects. A signature is generally a property of a class. Creating signatures can be thought of as a type of semantic activity, but applying signatures is a type of observation or judgment activity. Signatures can be stated as axioms or, perhaps more commonly, rules of judgment. Every investigation should consider the processes for suggesting, determining, and using signatures.

Data Collection Activities – Data Collection Activities are processes used to acquire data. This includes processes for collecting primary source data or secondary source data. Primary sources are data from collection processes which directly measure the environment or subjects of investigation. An example of a primary source data collection activity is an electro-optical satellite image acquisition processing chain or a park police officer interrogation of an arrested poacher. Secondary source data are data not created from a primary source, typically created by *agent*s and activities outside the direct influence of the investigation. Two examples of secondary source data are twitter messages pertaining to a specific subject and observations from an unrelated investigation, such as a neighboring country's list of poaching suspects. Primary sources are referred to simply as sources in this document. Secondary sources are usually observations and judgments from other investigations, where the data models and methods used to create the data might not be known.

Observation Activities – Observation Activities are the processes used for creating observations. An example of a human-based observation activity is the visual identification of a specific elephant made during a field study. The observation activity would specify the procedures for conducting a transect and recording the elephant sighting. An example of a machine-based observation activity is an unsupervised remote sensing classification of ecotypes which creates a polygon representing the geographic extent of an instance of an ecotype. Observation activities are

typically entity extraction procedures which use primary source data and the signature properties within the investigation's domain ontologies to identify instances of such classes and their observed property values.

Judgment Activities – Judgment Activities are the processes used to create judgments. The distinction between an observation and a judgment is described later in the document in detail. Most simply stated, the input required to create a judgment is not constrained to a primary source and the signatures of a class of objects, whereas an observation is so constrained. An example of a judgment activity is a software application that predicts the location of the next elephant poaching event. There is no such agreed signature of observational patterns that would suggest the next poaching location with so little uncertainty it is considered fact. Yet, the output of such a method might be useful in an investigation. A judgment activity is an analytic procedure used to produce assertions using all relevant data and for which no agreed-upon signature exists within the investigation.

Activity Activities – activities themselves must be reified. Activity Activities are processes used to create activities. For example, a software engineer will write code to create a new analytic procedure (*analytic activity*). The undertaking, or activity, of writing the code has attributes and metadata itself, such as how long the application took to develop and the necessary development tools.

Activity Set – Activity Set is a group of activities related for a common purpose. It may be useful to link all reified activities used to locate poachers, for example.

Dependent Elements – an activity is dependent on another activity and an *agent*. There must be an activity that created the activity. Often, two activities generate a new activity: 1) a previous version of a process, and 2) the activity that generated the new version. There must also be an *agent* responsible for each activity.

Activity
what it is: physical and logical processes *what it does*: enables agents to learn
key ideas: machine or human methods and workflows required to derive an instance of any element requires an agent to invoke

Figure 8: Activity element summary

3.2.5 Source

Source refers to the data created from a direct sensing of the environment or subject. This is the most primitive form of data pertaining to an investigation. Often source data will simply measure a dimension constrained by a space, or "field of view", over a period of time. There is little or no meaning inherent to source data. For example, a satellite image might be comprised of a matrix of numeric values ranging from 0-255 representing visible light measurements of part of the earth surface for one second duration. No meaning related to the problem domain is inherent to the pixels. Primary source data can be equated to the stimulus in the SSN ontology. Only through the application of observation activities to the pixel data is information about a problem domain related to the pixels. The concept of source data is analogous to the output of human senses. The eyes and ears transmit energy of specific dimensions from the surrounding environment, and the brain interprets the data, ascribing meaning to what is sensed.

Purpose – Source data is the most objective representation of a subject. Unlike observations and judgments, which ascribe meaning to source data through the applications of methods and domain models, source data is free of domain assumptions. Differentiating source data from observations and judgments is important because it allows for various domain models and observation activities - with different perspectives, assumptions, and knowledge goals – to be applied to the same source data. It is recognized that the objective qualities of source data is not easily achieved with humans serving as the sensor due to the inability to represent sensing and observing as having discrete outputs from the human mind. The O&M referenced in both the ODOE and ACRID frameworks, does not sufficiently recognize the distinction between sensors and the primary source data created by sensors. It is true that some sensors measure specific properties of specific objects, thereby generating observations. However, this is not the predominant case. More often, sensors create data which are interpreted by humans or machines for a particular domain model. Since reasonable differences can exist between domain models, it is important to differentiate the measurement function of a sensor from the activity that relates the measurement to an object in a domain model. For example, an acoustic sensor might record acoustic events of a specific frequency and amplitude range. The Hertz

and decibel values recorded are source data. To create an observation, an observation activity relates the acoustic measurements recorded to potential objects of interest, such as an elephant distress call or a gunshot.

Design Considerations – there are several source themes. The first theme is distinguishing between human and machine generated source data. The second set of themes involves the degree of abstraction of source data from the dimensional measurements of the sensor(s), such as between primary source data and secondary source data (or further degrees of separation from the sensor) and the consideration of processing chains. The third set of themes are considerations for the properties of source data. For example, there may be specific attribute and metadata standards for different types of sensor data. Finally, the physical sources, which create or store source data are sometimes important to represent.

Human and Machine Sensors data – as mentioned above, it is usually not possible to distinguish a person's description of what they see, hear, feel, smell, or taste (their senses) from the basic analysis and interpretation of what they sense. For example, imagine describing the smell of peach pie. Abstracting the nerve stimulation in the nose from the brain's interpretation of the nerve stimulation is difficult if not impossible. Further, the brain tightly integrates the five senses – often subconsciously. Consider witnessing a car accident quantitatively and normalized across each sense. Because pure, objective measurements are rarely available from human sources, it is important to consider each person – vice the individual senses of each person - as a unique sensor, with qualities of perception and memory that can vary based on personal attributes and domain understanding. Machine-based sensors can vary in performance as well, but the output data has a purely logical representation of the dimension of a subject or scene. Therefore, class properties of machine-based sensors are more stable and hereditary to subclasses.

Abstraction from sensor – source data are not always direct measures of the environment. When primary sources are used, there is provenance to a direct measure of the environment. But this is not always possible. Investigations often must use data already created from sources not under the custody of the investigation and sources that do not have provenance attribution to primary sources. These sources are referred to as secondary sources. An example is the use of a vector map dataset created from an unrelated investigation and lacks attribution relating the map content to sensor data. Investigations which distinguish data born via primary sources from data born via secondary sources will identify where observation-source provenance is broken. This can help determine the reliability of observations and judgments. Often, sensor data is used to create observations only after a processing step that transforms sensor data into a format that is useful for observation activities. It is important to consider that such processes might alter the data property values generated by the sensor. For example, if a Light Detection and Ranging (LiDAR) image of a farm is created by a sensor, and then it is projected to a local earth projection in a processing phase, it is possible that the projection process displaces some points in the LiDAR point cloud. Another processing step might identify only those elements of the point cloud associated with vegetation. Thus it is important to examine the collection and processing methods for each source to determine the source data processing activities which aught be accounted for and reified as a *data collection activity*.

Complex Sources – some observations can only be made if more than one source, or a stream of source data, are interpreted together. For example, interpretation of synthetic aperture radar (SAR) imagery can detect changes to the reflectance of a surface by comparing two images of the surface taken at different times. This is done by creating a composite image representing the differences between the original images. An observation recording a surface change should cite the composite image as its source. It is also increasingly common to have an observation of an object possible only through a method which interprets more than one stream of sensor data. Therefore, source data management must usually account for tracking source data processing chains so that observations can relate to the correct set of source data.

Source Properties – source data have varying format, structure, dimensions measured, and richness. Consideration of the properties of sources used in an investigation can improve how source data are managed and used in observation activities. The observation activities and resulting observation content from a real-time streaming video source will likely vary widely from the observation of a static photograph. The video feed likely has richer data and is continuous, with fleeting opportunities to create observations. One would expect streaming video interpretation to employ a much more complex set of activities for creating observations than the activities for interpreting the content of a static image. Likewise, data collection domains vary in the type and precision of metadata available for the acquired source data. For example, the geospatial precision of some collection methods may be within a few meters, other methods might be precise to a kilometer. And some data collections activities might not always provide geospatial data.

Physical Source – objects that store source data can be considered sources. For example, the hardware that a database resides on are physical sources. Likewise, a book that contains the field measurements is also a physical source. It can be important to represent the physical sources as a digital entity, and then relate observations to the physical sources to which they are associated.

Source Sets – there is often a need to identify sets of sources that are used for a common purpose or to which specific procedures shall apply.

Dependent Elements – source content is dependent upon Activities and other sources. Sources are created from an activity that acquires the source. The activity to acquire the source will use at least one physical source (sensor) and perhaps other source data.

Source
what it is: data measuring a dimension of the environment *what it does*: enables observation
key ideas: provides sensing of reality free of problem domain bias created by human (witness) or sensor collection activity secondary sources lack provenance to direct measurement

Figure 9: Source element summary

3.2.6 Observation

An observation is the instance of an information result created by an activity, which discovers instances of subclasses using ontology signatures applied to specific source data. Observing is accomplished through the application of an observation activity (a means of observing) to data from a source (such as an electro-optical satellite image). Recall that observation activities identify the instances of objects or object properties using, at a minimum, the signatures of such objects and object properties and a match of those signatures within primary source data.

The Observation element in this model uses the observation concepts of both the O&M and SSN models. According to the OM standard, an observation is "an event that estimates the observed property of some feature of interest [object] using a specified procedure and generates a result." We also incorporate the distinction of sensor data from sensors as described in the SSN model. Essentially, an observation provides 'factual' meaning to primary source data. It can be thought of as the first step in creating information from data.

Purpose – it is important for investigations to differentiate agreed upon facts from judgments, which rely on subjective analytic methods or sources without clear pedigree. Observations serve as the set of relevant facts as agreed upon by the agents of an investigation and as evidenced by primary source data. Differentiating observation from judgment allows investigations to present assertions that are without argument among the agents of an investigation from the assertions that are not based on an agreed upon object or property signature.

Design Considerations – there are three content themes specific to observations, which require consideration in the design of systems. Observations can be simple – not dependent on other observations - or complex. A more subtle theme is defining a difference between source data and observations. Finally, not observing something can be as important as observing it. This is referred to as negation. Determining how to model the absence of something is an important consideration. These three themes are described below.

Simple and Complex Observations – the complexity of an object or object property signature determines the complexity of an observation of that signature. Complex observations require the existence of at least one other observation and an inference rule, simple observations have no such dependence. The degree of complexity of an observation can be measured by the number of conditions, or premises, required for its associated signature. For example, a pachyderm ontology might specify distinct classes and signatures for elephant, elephant family, and elephant herd. Observing the instance of an elephant, on a single source, such as an air photo, by a human interpretation process is a simple observation. It does not require another observation to be valid. Observing a family of elephants, on the other hand, might require first observing several individual, related elephants in a particular pattern. An observation is valid, however complex, so long as all premises of the signature are true.

Abstraction from sensor data – data comprises an observation when an instance of an object class or object property is inferred from primary source data. Primary source data are often transformed or processed in some way that enables observation activities to be more easily or efficiently applied. Some sensors are unambiguous about the objects and properties they measure – essentially outputting observations. For example, a barometers measure atmospheric pressure at its location. If its location is known, then the observation of air pressure at that location is not typically distinguished from the sensor data of the barometer – at least for most domain problems.

Negation – it is often necessary for an investigation to note the absence of observing an object or object property. For example, if park police are searching house-to-house for a poacher, it is important to report a poacher

"not present" at every house they do not find the poacher. With sources that have broad coverage, defining something as not present might require an observation activity that uses the extent of the source to bound the range within which an object is not present.

Dependent Elements – observations are dependent upon source data, observation activities, and a domain ontology. Observations are meaningful information about an object or object class that is derived from applying an observation activity to primary source data. Objects and object classes are described in an ontology. Objects can have explicit or implicit signature properties.

Observation
what it is: identification of an object or property from source *what it does*: links sensor data to relevant meaning
key ideas: matches source data to an object 'signature' within domain ontology complex observations depend on existence of other observations provides lineage from judgments to primary sources

Figure 10: Observation element summary

3.2.7 Judgment

A judgment is an analytic assertion not directly observed. By contrast, observations are analytic assertions observed through the application of signatures to primary sources. Judgments should have direct or indirect links to all evidence. Here, the term evidence is used to reference the set of data used to determine a judgment. Evidence can include a set of observations, a set of judgments, and the set of analytic activities used to create the judgment.

Purpose – judgments provide content about an ontology or the instances within an ontology which are not directly observable. While observations are important for stating what is empirically true, the goals of most investigations are the output of higher level conclusions forming the synthesis of what is observed.

Design Considerations – judgments are used to express three types of ideas: hypothesis, models, and analytic judgments. These three types of judgment are described below.

Hypothesis. The first judgment type described is the hypothesis. A hypothesis is a supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation. Investigations often start with a hypothesis and a goal of accepting, rejecting, or refining the hypothesis. Therefore the outcome of an investigation with scientific applications is typically a refined hypothesis, or set of hypothesis, regarding a domain.

Semantic Judgment – making changes to a domain model requires making a judgments about how to abstract a real world phenomena. This requires considerations for the purpose of an investigation, agent domain knowledge, and the costs and benefits associated with more or less model precision. Changes to a domain model are considered a type of judgment distinct from hypothesis and analytic assessments. For example, part of a domain model might aggregate several object classes into a single class because the return on investment associated with creating instances for a precise set of classes is less than the return on investment for a single, simple class. Semantic judgments might have unique properties that link to evidence for modeling decisions.

Analytic Judgments – judgments created about the instances of objects within a domain are referred to as analytic judgments. It is important to realize that possible outcomes from analysis in an investigation include new conceptualizations, new relationships, new links to other domains, or predictions of future conditions that were not formalized prima facie. This has two implications: 1) analytic judgments can generally have a more complex information with a less structured data format and 2) analytic judgments typically precede semantic judgments. The use of analytic judgments can be helpful in expressing evidence-based opinions, which facilitate the debate and refinement of the domain model.

Dependent Elements – judgments are dependent only upon the activity element. While judgments do need an activity that creates the judgment, they do not need any supporting evidence to be valid. For example, an analyst might make a prediction of an event based on intuition in the absence of supporting judgements or observations.

Judgment
what it is: statement of belief absent direct observation *what it does*: differentiates observed 'fact' from analytic conclusion
key ideas: result of analysis of sources, observations and other judgments do not need evidence to be valid, unlike observations hypothesis and modeling decisions are forms of judgment

Figure 11: Judgment element summary

3.3 Implementation Paths

There are several models for implementing many elements within the investigation design. The SSN and O&M standards provide a good basis for modeling *observation* data and *source* data. Observations reference objects defined by an ontology. Also, there are several resources for creating, storing, and sharing ontologies, such as Protégé and Topbraid Composer. Ontology management however is outside of the scope of this paper.

Creating data requires activities that apply processes to data to create new data. *Activities* have always been at work in knowledge acquisition. The challenge is reifying key activities so that data results can be traced to the methods from which they were born. Activities can be reified through provenance metadata creation. Though much work can be done to enable provenance metadata recording in most data analysis software, the PROV Recommendation from W3C provides a conceptual model for provenance metadata. Tools like VisTrails enable the capture of provenance data in workflows.

There are several semantic models for describing *agent*s and their roles. The Friend of a Friend (foaf) vocabulary describes general attributes of people (Graves, Constabaris, & Brickley, 2007). The PROV ontology uses the foaf vocabulary with few additional properties for the people and organization subclasses of *agent* (Missier et al., 2013). The W3C Organization Ontology provides a basic model for describing organizations and relating people to organizations (W3C, 2015.). The PAV ontology began to make important distinctions in the roles of *agent*s for creating data in the biological sciences. It distinguishes authors, curators, and users (Ciccarese et al., 2013). However, more specific roles beyond those defined in the semantic vocabularies known to the author are needed for complex and collaborative knowledge organizations.

The two concepts which are not fully developed or addressed in literature are the *Judgment* and *Directive* elements. Future work will focus on incorporating these concepts into knowledge acquisition workflows. There are several reasons semantic web technologies and Linked Data principles are well suited for Directive design and implementation. First, there is a fundamental need to share data within and among investigations. This is the purpose behind the semantic web. Second, collaborations typically develop organically through self-organizing networks. Agents of separate investigations join forces in some capacity only after they discover one another. So investigation data must be available to query and return meaningful results. Third, reifying a Directive with a universal resource identifier (URI) can provide a single point of access to all related information about the investigation. The URIs to its component elements allows investigations to share select artifacts. For example, one investigation might choose to link to (and reuse) the observation procedures of another investigation that is using the same sensor data but for a different domain problem. Linking helps to facilitate the natural organization of investigations and the specific data elements within them.

4 ILLUSTRATIVE EXAMPLE

The running example used to discuss investigation design is summarized in this section. Three illustrations show how the information elements relate to one another and provide the necessary structure for supporting collaborative investigations. Figures 12 and 13 represent the recursive nature and interconnectedness of investigations. Figure 12 shows the relationship among the highest level directives in the example. Figure 13 shows functional directives born from these high level directives through additional planning activities, which often consider multiple parent directives. Investigations are executed through a common pattern of collecting and analyzing data within the constraints of their directives. Figure 14 shows how the elements represent such data in the given example.

Knowledge is often needed to achieve specific real outcomes. In the running example of elephant conservation, the country president initiates a series of related investigations by creating a *Directive 1.0*, the *Elephant Population Increase* directive. The president assigns the Minister of National Parks to lead this national effort intending to grow the wild elephant population. In turn, the Minister of National Parks creates two new investigations with Directive

1.1, one headed by the Chief of Park Police and another headed by the Chief Wildlife Biologist. Bear in mind that these figures represent directives and thus do not detail all the properties a directive might have, such as spatial or temporal boundaries, mandated *agent* relationships, or data standards.

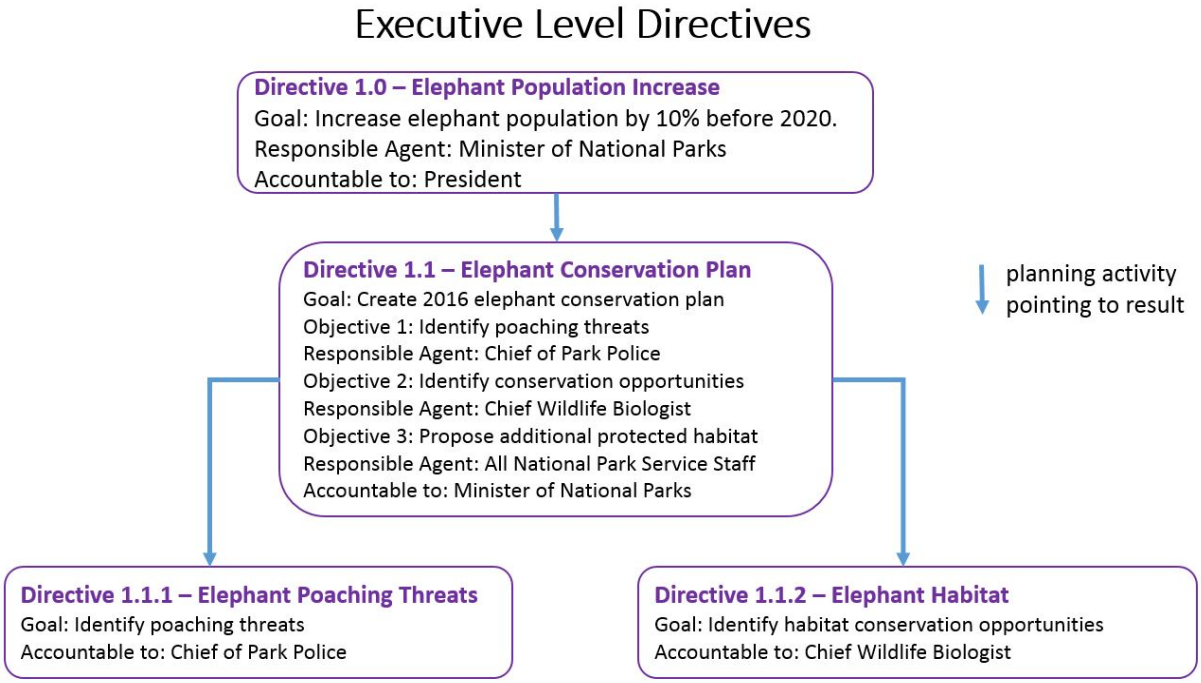

Figure 12. Executive level directives initiating the elephant conservation investigation.

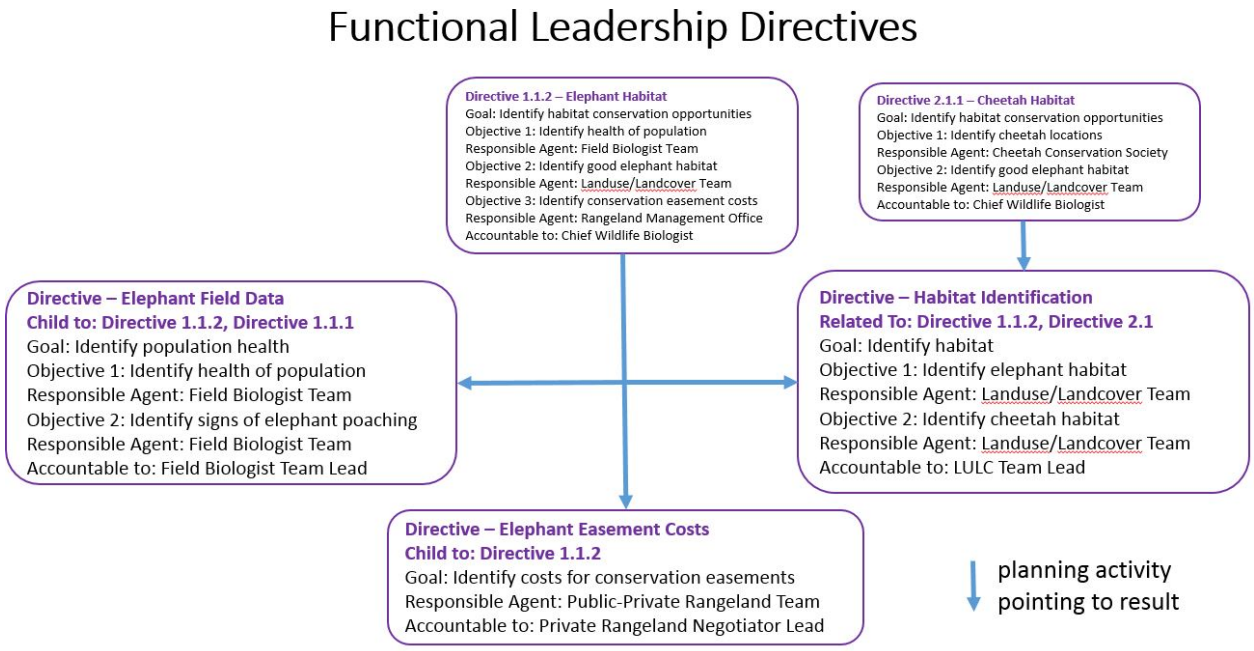

Figure 13. Functional level directives organize activities for all relevant investigations for which an *agent* is responsible.

The Elephant Habitat Investigation is led by the Chief Wildlife Biologist. Her office directs the Field Biologist Team to collect field data that would assess the health of the population and conservation opportunities through direct field observation. She also directs the Land Use Land Cover (LULC) team to identify elephant ecosystem

habitat across a wide geographic area. The LULC Team Lead crafts an investigation for his team that incorporates the elephant habitat request with a similar directive requesting the identification of cheetah habitat for the same geography.

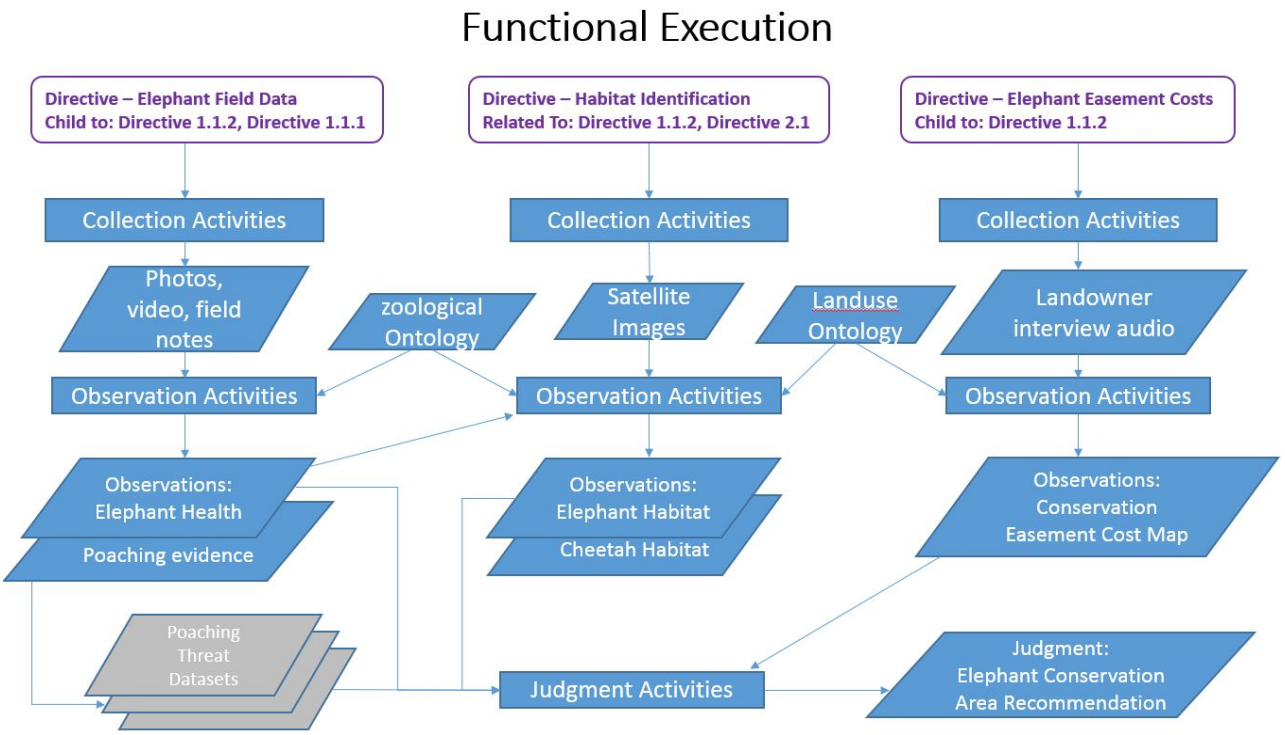

Figure 14. Functional execution includes the lower level activities that generate knowledge.

Finally, three directives guide the three teams of specialized experts in collecting sources, making observations, and deriving judgments about how best to improve elephant conservation. These three directives inherit the constraints imposed by the higher level directives, such as a controlled vocabulary or observation data standards to ensure interoperability for judgment activities.

Each team directive might also convey workflow constraints tailored to the specific objectives and expertise of the team. The LULC Team, for example, might define specific collection activities to acquire satellite imagery of the national parks and adjacent areas. These collection activities will vary greatly from the *in situ* data collection activities of the Field Biologist Team.

What is common across all teams is the pattern of data maturity. The most basic task is to establish what information is needed and the rules for its creation. This is represented by the directive. For empirical studies, a pure measure of the subject is the most desirable, primary data input to the workflow. For the LULC team, this source would be satellite imagery. Observations relate sets of pixels in the imagery dataset to object classes in the domain ontologies identified by the directive. Likewise, photos and field notes or landowner interview recordings provide source data from which observations are born for the other two investigations. Observations and existing judgments from all available sources within and external to these investigations are used in the judgment activity. Judgment activity represents the processes used to create a judgment regarding elephant conservation. It is important to note that several judgments could be made, some of them conflicting. With provenance to the *agent*s, information, and methods used to derive each judgment, decision makers can evaluate the evidence supporting each judgment and decide which, if any, meet their information needs.

This investigation design presents a workflow pattern common to all knowledge-seeking efforts. While many of the elements are not formally modeled, or reified, in most contemporary investigations, it is argued here that this must change to achieve an integrated knowledge-seeking enterprise. Representing investigations as a package of elements guided by a directive will improve execution, engender trust, and enable repeatability of investigations. The value of this design extends to external efficiencies. Investigators should be able to “copy and paste” another investigation of a similar problem domain – in its entirety – to use the data directly or make workflow changes

specific to their unique circumstances and re-implement – bypassing much of the cost of workflow creation. Researchers will search for complete workflows and data, not simply for datasets or specific methods. For example, a different team of researchers might be mapping elephant habitat in East Africa. They might search, find, and recycle the Elephant Habitat investigation with minor changes specific for the East African environment. In fact, a convention for investigation design, such as the one espoused here, has the potential to stimulate an economic market for workflow exchange.

5 CONCLUSION

We are experiencing an era of big data, specialist sources and methods, and abundant access to data and knowledge reporting tools for the masses. However, while collaboration exists in highly visible efforts with a dedicated following and respective tool support, what has inhibited generla widespread collaboration across arbitrary efforts is the *lack of a common design pattern for recording what is learned and how it is learned in the course of research.* To move beyond the sharing of datasets to the actual sharing of knowledge, one needs to formalize and document information, which can be made explicit about the entire knowledge acquisition effort, i.e., its purpose, structure, and findings. Further, a lack of standardized requests for knowledge acquisition activity results in multi-party efforts with multiple interpretations of a request. Establishing a convention for a common investigation design allows for specific results and for an investigations to be reusable and therefore more valuable.

We introduce *investigation design* (ID) as a meta model designed to meet three goals: (1) account for all data needed to create, use, and recreate an investigation, (2) be relevant to any domain, and (3) be relevant to any administrative level of an organization – requesting, managing, or creating domain knowledge. ID identifies an investigation as the collection of data elements for a knowledge-seeking effort. *Source*, *Observation*, and *Judgment* are content elements used within an investigation to refine a domain ontology. In fact, one can consider a domain ontology simply as the current set of accepted ontological judgments about a set of related objects. *Agent* and *Activity* are content elements. Activities are physical or algorithmic processes used to generate data, e.g., evaluating an image by means of an algorithm or expert. *Agent*s are the people responsible for activities in an investigation. Instances within each content element should be linked to their dependent elements through provenance attribution.

Future work will expand on the subclasses and properties of directives to arrive at a clear expression of goals and constraints of an investigation. Given the complex relationships among investigations, often fortuitously linked ex post facto, a linked data approach is a candidate implementation for directives. Judgments are assessments requiring more than just source data as evidence, thus how judgments are represented distinct from observations is the second focus of future efforts. Finally, relating the data principles to those of organizational development and collective action research is ultimately needed to successfully transfer this model into practice.

References

Abels, S., Ahlemann, F., Hahn, A., Hausmann, K., & Strickmann, J. (2006). PROMONT – A Project Management Ontology as a Reference for Virtual Project Organizations. In R. Meersman, Z. Tari, & P. Herrero (Eds.), *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops* (pp. 813–823). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/11915034_105

Bose, R., & Frew, J. (2005). Lineage retrieval for scientific data processing: a survey. *ACM Computing Surveys (CSUR)*, *37*(1), 1–28. http://doi.org/10.1145/1057977.1057978

Bose, R., & Frew, J. (2005). Lineage Retrieval for Scientific Data Processing: A Survey. *ACM Computing Surveys*, *37*(1), 1–28.

Botts, M., Percivall, G., Reed, C., & Davidson, J. (2006). OGC® Sensor Web Enablement: Overview and High Level Architecture (pp. 175–190). Presented at the International conference on GeoSensor Networks, Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-79996-2_10

Bröring, A., Echterhoff, J., Jirka, S., Simonis, I., Everding, T., Stasch, C., … Lemmens, R. (2011). New Generation Sensor Web Enablement. *Sensors*, *11*(3), 2652–2699. http://doi.org/10.3390/s110302652

Car, N. J. (2013a). A method and example system for managing provenance information in a heterogeneous process environment - a provenance architecture containing the Provenance Management System (PROMS). Presented at the 20th International Congress on Modelling and Simulation, Adelaide, Australia. Retrieved from http://www.mssanz.org.au.previewdns.com/modsim2013/C7/car.pdf

Car, N. J. (2013b). A method and example system for managing provenance information in heterogeneous process environment - a provenance architecture containing the Provenance Management System (PROMS). In *20th International Congress on Modelling and Simulation, Adelaide, Australia, 1-6 December 2013*. Adelaide, Australia.

Compton, M., Barnaghi, P., Bermudez, L., García-Castro, R., Corcho, O., Cox, S., … Taylor, K. (2012). The SSN ontology of the W3C semantic sensor network incubator group. *Web Semantics: Science, Services and Agents on the World Wide Web*, *17*, 25–32. http://doi.org/10.1016/j.websem.2012.05.003

Cox, S. (2013). OGC Abstract Specification, Geographic information - Observations and measurements. OGC. Retrieved from http://www.opengix.net/doc/is/om/2.0

Davidson, S. B., & Freire, J. (2008). Provenance and Scientific Workflows: Challenges and Opportunities. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data* (pp. 1345–1350). New York, NY, USA: ACM. http://doi.org/10.1145/1376616.1376772

de Moor, A., De Leenheer, P., & Meersman, R. (2006). DOGMA-MESS: A Meaning Evolution Support System for Interorganizational Ontology Engineering. In H. Schärfe, P. Hitzler, & P. Øhrstrøm (Eds.), *Conceptual Structures: Inspiration and Application* (Vol. 4068, pp. 189–202). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://link.springer.com/10.1007/11787181_14

Elrakaiby, Y., Cuppens, F., & Cuppens-Boulahia, N. (2012). Formal enforcement and management of obligation policies. *Data & Knowledge Engineering*, *71*(1), 127–147. http://doi.org/10.1016/j.datak.2011.09.001

Freire, J., Koop, D., Santos, E., & Silva, C. T. (2008a). Provenance for Computational Tasks: A Survey. *Computing in Science & Engineering*, *10*(3), 11–21. http://doi.org/10.1109/MCSE.2008.79

Freire, J., Koop, D., Santos, E., & Silva, C. T. (2008b). Provenance for Computational Tasks: A Survey. *Computing in Science and Engg.*, *10*(3), 11–21. http://doi.org/10.1109/MCSE.2008.79

Freire, J., Silva, C. T., Callahan, S. P., Santos, E., Scheidegger, C. E., & Vo, H. T. (2006a). Managing Rapidly-Evolving Scientific Workflows (pp. 10–18). Presented at the International Provenance and Annotation Workshop, Springer Berlin Heidelberg. http://doi.org/10.1007/11890850_2

Freire, J., Silva, C. T., Callahan, S. P., Santos, E., Scheidegger, C. E., & Vo, H. T. (2006b). Managing Rapidly-evolving Scientific Workflows. In *Proceedings of the 2006 International Conference on Provenance and Annotation of Data* (pp. 10–18). Berlin, Heidelberg: Springer-Verlag. http://doi.org/10.1007/11890850_2

Graves, M., Constabaris, A., & Brickley, D. (2007). FOAF: Connecting People on the Semantic Web. *Cataloging & Classification Quarterly*, *43*(3-4), 191–202. http://doi.org/10.1300/J104v43n03_10

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, *43*(5–6), 907–928. http://doi.org/10.1006/ijhc.1995.1081

Harth, A., & Gil, Y. (2014). Geospatial data integration with linked data and provenance tracking. In *W3C/OGC Linking Geospatial Data Workshop*. Retrieved from http://www.isi.edu/~gil/papers/harth-gil-lgd14.pdf

Heath, T., & Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, *1*(1), 1–136. http://doi.org/10.2200/S00334ED1V01Y201102WBE001

Kandel, E. R., Markram, H., Matthews, P. M., Yuste, R., & Koch, C. (2013). Neuroscience thinks big (and collaboratively). *Nature Reviews Neuroscience*, *14*(9), 659–664.

Lakshmanan, G. T., Curbera, F., Freire, J., & Sheth, A. (2011). Provenance in web applications. *IEEE Internet Computing*, *15*(1), 0017–21.

Laney, D. (2001). 3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety. META Group Research. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Li, S., Dragicevic, S., Veenendaal, B., & Brovelli, M. A. (2013). Theme section “Towards Intelligent Geoprocessing on the Web.” *ISPRS Journal of Photogrammetry and Remote Sensing*, *Complete*(83), 138–139. http://doi.org/10.1016/j.isprsjprs.2013.07.007

McKee, L. (2015). OGC Information Technology Standards for Sustainable Development. OGC. Retrieved from www.opengeospatial.org/docs/whitepapers

Missier, P., Belhajjame, K., & Cheney, J. (2013). The W3C PROV Family of Specifications for Modelling Provenance Metadata. In *Proceedings of the 16th International Conference on Extending Database Technology* (pp. 773–776). New York, NY, USA: ACM. http://doi.org/10.1145/2452376.2452478

National Academies. (2004). *Facilitating Interdisciplinary Research*. Washington, DC, USA: National Academies Press. Retrieved from http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10078093

Nooteboom, B. (2000). Learning by interaction: absorptive capacity, cognitive distance and governance. *Journal of Management and Governance*, *4*(1-2), 69–92.

OGC Testbed 10 Provenance Engineering Report. (2014). OGC. Retrieved from http://www.opengeospatial.org/docs/er

Pomponio, L., & Le Goc, M. (2014). Reducing the gap between experts’ knowledge and data: The TOM4D methodology. *Data & Knowledge Engineering*, *94, Part A*, 1–37. http://doi.org/10.1016/j.datak.2014.07.006

Project Management Institute. (2008). *A GUIDE TO THE PROJECT MANAGEMENT BODY OF KNOWLEDGE (PMBOK® Guide)* (Fourth Edition). Project Management Institute. Retrieved from http://proquest.safaribooksonline.com.mutex.gmu.edu/9781933890517

Schreiber, G. (2000). *Knowledge Engineering and Management : The CommonKADS Methodology*. Cambridge, Mass: The MIT Press. Retrieved from http://mutex.gmu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=21785&site=ehost-live&scope=site

Shaon, A., Callaghan, S., Lawrence, B., Matthews, B., Osborn, T., Harpham, C., & Woolf, A. (2012). Opening Up Climate Research: A Linked Data Approach to Publishing Data Provenance. *International Journal of Digital Curation*, *7*(1), 163–173. http://doi.org/10.2218/ijdc.v7i1.223

Silvertown, J. (2009). A new dawn for citizen science. *Trends in Ecology & Evolution*, *24*(9), 467–471. http://doi.org/10.1016/j.tree.2009.03.017

Spyns, P., Meersman, R., & Jarrar, M. (2002). Data modelling versus ontology engineering. *ACM SIGMod Record*, *31*(4), 12–17.

Staab, S., & Studer, R. (2004). *Handbook on ontologies*. Berlin ; New York: Springer.

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, *25*(1–2), 161–197. http://doi.org/10.1016/S0169-023X(97)00056-6

van Rijnsoever, F. J., & Hessels, L. K. (2011). Factors associated with disciplinary and interdisciplinary research collaboration. *Research Policy*, *40*(3), 463–472. http://doi.org/10.1016/j.respol.2010.11.001

W3C. (n.d.). The Organization Ontology. Retrieved October 21, 2015, from http://www.w3.org/TR/vocab-org/