Ontology engineering in provenance enablement for the National Climate Assessment

Xiaogang Ma \(^a\)^*, Jin Guang Zheng \(^a\), Justin C. Goldstein \(^b,c\), Stephan Zednik \(^a\), Linyun Fu \(^a\), Brian Duggan \(^b,c\), Steven M. Aulenbach \(^b,c\), Patrick West \(^a\), Curt Tilmes \(^c,d\), Peter Fox \(^a\)

\(^a\) Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, USA
\(^b\) University Corporation for Atmospheric Research, 3300 Mitchell Lane, Boulder, CO 80301, USA
\(^c\) U.S. Global Change Research Program, Suite 250, 1717 Pennsylvania Avenue, NW, Washington, DC 20006, USA
\(^d\) NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD 20771, USA

* Corresponding author. Tel: +1 518 276 4384; fax: +1 518 276 4464.
Email addresses: max7@rpi.edu (X. Ma), zhengj3@rpi.edu (J.G. Zheng), jgoldstein@usgcrp.gov (J.C. Goldstein), zednis2@rpi.edu (S. Zednik), ful2@rpi.edu (L. Fu), bduggan@usgcrp.gov (B. Duggan), saulenbach@usgcrp.gov (S.M. Aulenbach), westp@rpi.edu (P. West), curt.tilmes@nasa.gov (C. Tilmes), foxp@rpi.edu (P. Fox)

Abstract:
The National Climate Assessment of the U.S. Global Change Research Program (USGCRP) analyzes and presents the impacts of climate change on the United States. The provenance information in the assessment is important because the assessment findings are of great public and academic concern and are used in policy and decision-making. By applying a use case-driven iterative methodology, we developed information models and ontology to represent the content structure of the recent National Climate Assessment draft report and its associated provenance information. The provenance information in our work was maintained in the context of the “Web of Data”. In addition to the pilot systems we developed, other tools and services are also able to retrieve and utilize the provenance information. Our work is part of a Global Change
Information System coordinated by the USGCRP that will eventually cover provenance information for the entire scope of global change research.

Keywords: Provenance, Ontology engineering, Use cases, Global change, Semantic Web

1 Introduction

Mandated by the U.S. Congress through the Global Change Research Act of 1990 (U.S. Code, 1990), the U.S. Global Change Research Program (USGCRP)\(^1\) coordinates the production of the quadrennial National Climate Assessment (NCA) report. As a federal assessment, the report is considered an authoritative resource for understanding and communicating climate change science and impacts in the United States\(^2\). Global change science builds on a huge collection of scientific research across the world, which also generates provenance information (Moreau and Missier, 2013) about entities, activities, people and organizations involved in the production of the research findings and the supporting datasets and methods. The NCA will build a traceable account as an assembled view of the provenance for each research finding in the NCA report. Capturing and presenting provenance of global change research, and linking to the publications, datasets, instruments, models, algorithms and workflows that support key research findings increases understanding, credibility and trust in the NCA works, and aids in fostering repeatability of results and conclusions (cf. Buneman et al., 2000; Frew and Dozier, 2012; Reichman et al., 2011; Simmhan et al. 2005).

Coincident with the production of the third NCA report (NCA3) that is scheduled for completion in early 2014 (USGCRP, 2013), the USGCRP is coordinating the development of a Global Change Information System (GCIS) that will present the content of that report and its provenance, including the scientific support for the findings of the assessment (Tilmes et al., 2013). The web-based GCIS provides a platform for seeking approaches to represent the provenance information and implementing the results with Semantic Web technologies. The Semantic Web (Berners-Lee et al., 2001; Hendler, 2003) is a “Web of Data” in addition to the

---


classic “Web of Documents”. The purpose of the “Web of Data” is not only to make data accessible on the Web but also to create links, to permit both humans and machines to read and explore the linked data available in the Semantic Web (Berners-Lee, 2006). To enhance the use of the “Web of Data”, semantic technologies leverage extant ontologies (Bauer and Kaltenböck, 2012; Hendler, 2003) that define important concepts and relationships in domains of relevance. Applications of featured functionalities can be developed based on those ontologies, such as mediation among heterogeneous data sources (Fox et al., 2009) and statistical analysis of gene product annotation data (Chan et al., 2012). For Semantic Web applications in the GCIS, a primary work is developing an information model and using/extending existing ontologies to represent the structure of content in NCA3 and the provenance information within it.

The purpose of this paper is to present and discuss our methods for developing and deploying Semantic Web technologies for the NCA, the resulting ontology, and its application in several pilot systems in the GCIS. Working with the publicly available draft version of NCA33 (released on 2013-01-11) we applied a use case-driven, iterative development method (Fox and McGuinness, 2008) to develop the GCIS information model. A use case is a flow of interactions between an actor and a system in order to achieve a specific objective (Bittner and Spence, 2003). In general use cases establish contexts in which scientists of a particular discipline (e.g., climate change, petrology, water resources, etc.) work with computer specialists on topics of interest in that domain. In our work we developed a number of use cases illustrating interactions between a variety of users (actors) and the desired GCIS. Through analyzing those use cases we defined objects and relationships in the conceptual and logical information models, and then associated them with classes and properties in a “GCIS ontology”. In organizing those classes and properties we referred to a number of existing ontologies from various domains. We enriched the GCIS ontology to include provenance information by incorporating the World Wide Web Consortium (W3C)’s recommendation on provenance modeling (PROV) (Groth and Moreau, 2013) into the design. Re-using and mapping to classes and properties of existing ontologies also reflects the central idea, “Web of Data”, of the Semantic Web. As illustrated by Berners-Lee (2006), the essential condition to turn data from four-star to five-star is linking the data to other

data (Table 1). Ontology re-using and mapping will link data in the GCIS to external data sources and thus improve the quality of the system in the Semantic Web.

Table 1 Star rating system for Linked Open Data (Berners-Lee, 2006)

| ★  | (1) Available on the web (whatever format) but with an open license, to be Open Data |
| ★★ | (2) Available as machine-readable structured data (e.g. excel instead of image scan of a table) |
| ★★★ | (3) As (2) plus non-proprietary format (e.g. CSV instead of excel) |
| ★★★★ | (4) All the above, plus: Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff |
| ★★★★★ | (5) All the above, plus: Link your data to other people’s data to provide context |

The use case-driven iterative method applied in this paper has its unique advantages for developing domain-specific Semantic Web applications, though it shares some features with existing methods for software and knowledge engineering, such as the PrIMe (Miles et al., 2011), the Content Ontology Design Patterns (Gangemi and Presutti, 2009) and the Agile Methods (Cohen et al., 2004). The ontology engineering work in this paper is one of the first to implement the W3C PROV Recommendation for open government data services. The abundant provenance information will increase the reliability and aid in repeatability of the results and findings presented by the GCIS. There are three stages for the global linked open data efforts: opening, linking and re-using (Bauer and Kaltenböck, 2012; Berners-Lee, 2006; Ding, 2012). GCIS will make content and associated information of NCA3 open and interlinked, and thus a part of the Web of Data (cf. Executive Order, 2013). Though the GCIS ontology was built with use case analyses surrounding NCA3, the ontology itself is reusable in other contexts for building provenance-aware data services. The procedure and examples discussed in this paper provide timely experience for re-using the W3C PROV and adapting the GCIS ontology in other applications.
The remainder of the paper is organized as follows. Section 2 introduces the development method applied for building the GCIS information models and ontology and some of the many use cases analyzed. Section 3 presents the resulting ontology and its application in provenance-enabled pilot systems. Section 4 compares our work with related works and discusses potential future extensions. Section 5 wraps up the paper with conclusions and discussion.

2 Materials and Methods

NCA3 was developed by 240 authors from academia, governmental agencies and private and non-profit sectors under the direction of the National Climate Assessment Development Advisory Committee (NCADAC), a U.S. Federal Advisory Committee serving to oversee the NCA activities in the USGCRP. On January 11, 2013, the NCADAC Draft Climate Assessment Report (hereafter as Draft NCA3) was released for public review and comments. The assessment integrates the findings of worldwide scientific observations and research of climate change, and analyzes the impacts of climate change on the United States. Topics in the resulting report include the evaluation and integration of the current scientific understanding of climate change, the related impacts on various societal and environmental sectors and regions of the U.S., and decision support and mitigation and adaptation strategies. Those topics are distributed in the 30 chapters and two appendices of the resulting Draft NCA3. The initial GCIS design and development, presented here, was based on the information publicly available in that draft report.

The content of Draft NCA3 mentioned above and the associated metadata, such as NCADAC members, chapter authors, chapter titles, and figures in each chapter, etc., provide the initial materials for us to recognize concepts and relationships in an information model, and define them as classes, instances and properties in the GCIS ontology. In particular, the information sources underpinning Draft NCA3 are important materials for use in tracing and representing provenance information. The draft report references approximately 3600 publications. Such abundant information sources provide a solid basis for provenance instances and the development of provenance components in the GCIS ontology.
2.1 An iterative method for ontology development and application

We applied a use case-driven development methodology (Fig. 1) to recognize and collect components that can be used to represent provenance information in assessments such as the NCA. This is a maturing method derived from a number of our previous works (e.g., Benedict et al., 2007; Fox et al., 2009; Rozell et al., 2010; West et al., 2012). The method is iterative and the key steps in a single round include drafting a use case, setting up a team, analyzing the use case and drafting an information model/ontology with tools, reviewing and refining the model/ontology, adopting a technical approach and building a rapid prototype application, and preparing for the next use case iteration. In this approach, evaluation and iteration can be applied to each step where appropriate (note inner arrows in Fig. 1). This approach also reflects the open-world aspect of the method, i.e. not all knowledge is known or encoded, and may be wrong, incomplete and/or evolving (Fox and McGuinness, 2008). Our experience shows that the method is easy to implement, and here we elaborate on the practical techniques in a few steps.

![Figure 1](image_url)

Fig. 1. An iterative development methodology for establishing and improving Semantic Web applications in a domain (from Fox and McGuinness, 2008).

Use cases are applied to identify goals/objectives to be accomplished, resources to be used to achieve these objectives, and methods to be used to produce the desired results. The goals are
generally proposed by domain experts/stakeholders. We adopted a use case template\(^4\) to collect information such as the objective that a primary actor wishes to accomplish, the preconditions of a system and the triggers to initiate the use case, the sequence of interactions between the primary actor and the system to achieve the objective, resources (e.g., datasets, models, and events, etc.) used in these interactions, and the post-conditions of the system. The template allows the domain scientists to describe understandable and operational use cases for the team members to discuss and analyze.

The iterative method is implemented by a small team with a mixture of skills and knowledge (Fox and McGuinness, 2008). A facilitator sets and monitors the direction of the iterative approach, provides guidance for scoping the use case and suggests a schedule of milestones for implementation. Another role is a scribe for documenting activities and decisions of the team, which may be performed concurrently by one of the team members. Team formation includes the design of roles within a team and recruitment of members. Roles within the team include domain experts or stakeholders with background knowledge of the overlying topic pertaining to a use case, data and information producers familiar with where to access and/or produce the resources required, knowledge and information modelers who will analyze components and processes in the use case and draw conceptual schemas for them, and software engineers who will collaborate with the modelers to leverage existing capabilities and develop prototype applications.

The approach advocates finding and using relevant tools. For example, as part of our distributed and collaborative development process we used Skype\(^5\) and GoToMeeting\(^6\) for organizing teleconferences with remotely-based participants, Titanpad\(^7\) for taking meeting notes, CMapTools\(^8\) for drawing conceptual maps, Notepad++\(^9\) and Protégé\(^10\) for editing and checking ontologies, Parrot\(^11\) for ontology documentation, Epimorphics Linked Data API

---


(ELDA)\(^{12}\) for developing a graphical user interface for browsing linked data, and S2S\(^{13}\) for developing a prototype of faceted data searching and browsing. Our intent is “from the community and for the community”, that is, we use open-source and free software programs as often as possible, and make results open-source and free to the community.

2.2 Provenance-explicit use cases

As our current work focuses on recognizing, representing and capturing provenance for Draft NCA3, we developed use cases surrounding that topic. Below are the briefs of three selected use cases.

\textit{Use Case 1}

- Title: Visit data center website of dataset used to generate a report figure
- Actor: A reader
- System: The GCIS website
- Flow of interactions: The reader wishes to identify the source of the data used to produce a particular figure in Draft NCA3. A reference to the paper in which the image contained in this figure was originally published appears in the figure caption. Clicking that reference displays a page of metadata information about the paper, including links to the datasets used in that paper. Pursuing each of those links presents a page of metadata information about the dataset, including a link back to the agency/data center web page describing the dataset in more detail and making the actual data available for order or download.

\textit{Use Case 2}

- Title: Identify roles of people in the generation of a chapter in Draft NCA3
- Actor: A reader
- System: The GCIS website


• Flow of interactions: A reader sees that Chapter 6 (Agriculture) in Draft NCA3 was written by a group of authors mentioned in a list. On the title page of that chapter the reader can view the role of each author, e.g., convening lead author, lead author or contributing author, in the generation of this report chapter.

Use Case 3
• Title: Provenance tracing of NASA contributions to Figure 1.2 in Draft NCA3
• Actor: A reader
• System: The GCIS website
• Flow of interactions: A reader sees that the caption of Figure 1.2 “Sea Level Rise: Past, Present and Future” of Draft NCA3 cites four data sources. Selecting the third citation displays a page of information about the cited paper and a citation to the dataset used in that paper. Information about the dataset includes a formal description of its origin, that is, the dataset is derived from data produced by the TOPEX/Poseidon and Jason altimeter missions funded by NASA (National Aeronautics and Space Administration) and CNES (Centre national d'études spatiales). Clicking a link to each of these missions presents a page about the platforms, instruments and sensors in that mission.

Use Case 1 was used to recognize the basic entities and agents to be included in GCIS and the GCIS ontology. The concept map describing relationships between concepts pertaining to Use Case 1 is presented as Figure 2. Concept maps are graphical tools for organizing and representing knowledge (Novak and Cañas, 2008). They are often used as the first step in information models that are pre-cursors to ontology engineering (Starr and de Oliveira, 2013). In Fig. 2a, instances are shown as rectangular boxes and their interrelationships as labeled arrows. In Fig. 2b, classes are shown as rounded boxes and properties as labeled arrows. We used a specific quadrilateral rectangle to display the classes, such as “xsd:string”. Because the model is for use in the Semantic Web, we distinguished two primary types of properties in it: object properties link individuals to individuals and datatype properties link individuals to data values. Examples of the former include “gcis:hasChapter”, “gcis:hasFigure” and “gcis:hasImage”, etc. and an example of the latter is “gcis:hasCaption”. In the analysis of Use Case 1 we used “gcis:Publication” since a figure can come from just about any publication types (Fig. 2b).
Moreover, we used an object property “gcis:sourceDataset” to describe the datasets underpinning both publications and figures.

Fig. 2. Analysis of Use Case 1 “Visit data center website of dataset used to generate a report figure”: (a) Information model fragment represented as an intuitive concept map of the use case and (b) Classes and properties recognized from the use case. For description of the use case see text in Section 2.2. The instances in (a) are labeled with generic names and do not directly applicable to instances in Draft NCA3 and other sources. Classes and properties defined in (b) are marked with light cyan color and a namespace prefix “gcis”. The prefix marks the source model/ontology of a class and the color makes the concept map easier for humans to read. The classes re-used from external ontologies are marked with a different color and namespace prefix, such as “dctype” (The Type Vocabulary in the Dublin Core Metadata Initiative (DCMI))\textsuperscript{14} and “org” (The Organization ontology)\textsuperscript{15}.

Ontology engineering differs from database schema modeling in that ontologies do not pertain to specific use cases (cf. Spyns et al., 2002). We recognize this distinction from two aspects: (1)

\textsuperscript{15} http://www.w3.org/TR/vocab-org/. Accessed on 2013-09-25.
that a defined class or property should be meaningful and robust enough to meet the requirements of various use cases, and (2) that an ontology can be extended by adding classes and properties recognized from new use cases through the iterative approach. Using “gcis:Publication” instead of “gcis:Paper” in the result of Use Case 1 reflects the former aspect. For the latter, we developed and analyzed more use cases to enrich the ontology, such as Use Case 2.

Fig. 3. The three primary classes in the PROV-O ontology and the properties that relate them (adapted from Lebo et al., 2013).

We decided to use the W3C PROV-O ontology (Lebo et al., 2013) to describe the role of each author in Use Case 2. PROV-O\textsuperscript{16} ontology provides a number of classes, properties and restrictions that are tailored for representing provenance information (Fig. 3). We defined objects in Use Case 2 as instances of corresponding PROV-O classes (Fig. 4). The eight authors in that chapter, three of which are shown in Fig. 4, each has a specific role (e.g., “convening lead author”) and a common role (i.e., “author”). Use Case 2 inspired us to identify roles in scientific works from other sources, such as the ISO 19115 and ISO 19115-2 standards (ISO, 2003; 2009)\textsuperscript{17}, to enrich the GCIS ontology. We mapped/linked a number of classes and properties in


the GCIS ontology into those in the PROV-O ontology. For example, we asserted “gcis:Person” as a sub-class of “prov:Person”, and “gcis:sourceDataset” as a sub-property of “prov:wasDerivedFrom”. By such mappings we can utilize reasoners (i.e., software tools able to complete logical inferences from a set of asserted facts) that are suitable for the PROV-O ontology, and thus to retrieve provenance graphs from the established GCIS, such as those shown in Fig. 4.

Fig. 4. Roles of selected authors in Chapter 6 Agriculture of Draft NCA3 described with PROV-O ontology: (a) Role of each author in Chapter 6 writing (i.e., the writing is an activity) described with “prov:qualifiedAssociation” and (b) Role of each author in Chapter 6 (i.e., the chapter itself is an entity) described with “prov:qualifiedAttribution”. Instances of classes “prov:Agent”, “prov:Activity” and “prov:Entity” are shown in orange, blue and yellow colors, respectively. “prov”, “rdf” and “dct” are the namespace prefix of the PROV-O ontology (footnote 16), the Resource Description Framework (RDF)\(^\text{18}\) and the DCMI Metadata Terms\(^\text{19}\), respectively. Instances are shown as rectangles, and the instance names are part of their corresponding uniform resource identifiers (URIs). For example, “person/Gene_Takle” follows a structure of “resource_type/resource_identifier”, and the full URI of it is “http://data.globalchange.gov/person/Gene_Takle”.

Use Cases 1 and 2 permitted the establishment of a basic conceptual structure of entities, agents and roles (e.g., report, chapter, figure, dataset, data center, and author, etc.) of Draft NCA3. In Use Case 3 we further analyzed activities and tackled the sensors, instruments, platforms, models, etc. from which datasets and publications are generated. By analyzing Use Case 3 we collected instances of calibration, model, software, sensor, instrument and platform. We also mapped the classes of those instances into corresponding classes in the PROV-O ontology, such as “prov:Agent”, “prov:Entity” and “prov:Activity” (Fig. 5).
Fig. 5. Concept map illustrating the component of NASA’s contribution to Figure 1.2 in Draft NCA3. Here only the details of one paper (i.e., “paper/103”) cited by that figure are shown: (a) Instances of calibration, model and software underpinning “paper/103” and (b) Instances of sensor, instrument and platform underpinning that paper. In (b) the details of source dataset “dataset/Jason1” and “dataset/Jason2” are omitted. Instances of classes “prov:Agent”, “prov:Activity” and “prov:Entity” are shown in orange, blue and yellow colors, respectively. Instances are shown as rectangles, and the instance names are part of their corresponding URIs.

2.3 Re-using and mapping to existing ontologies

In the above use case analyses we have demonstrated the re-use of a few existing ontologies, such as the DCMI Types Vocabulary (see footnote 14 and classes with a namespace prefix “dctype” in Fig. 2b), the Organization ontology (see footnote 15 and classes with a prefix “org” in Fig. 2b), and the PROV-O ontology (see footnote 16 and classes with a prefix “prov” in Figs. 4 and 5). Besides these, our work also involved a number of other ontologies, such as the Bibliographic ontology (with a prefix “bibo”)20, the FOAF (Friend of a Friend) ontology (with a prefix “foaf”)21, the SKOS (Simple Knowledge Organization System) model (with a prefix “skos”)22 and the Place ontology (with a prefix “place”)23, etc.

We use those ontologies based on two rationales: reduce duplicate works (especially ontologies) and promote interoperability. Most of those ontologies are formal recommendations based on group or community works and are commonly used in the Semantic Web. In our use case analyses, we discussed topics which overlapped with one or more existing ontologies within our team to see if we can adopt or adapt classes and properties in those ontologies. For example, in analyzing the sub-categories of agents, we referred to the sub-classes (i.e., “foaf:Organization”, “foaf:Group” and “foaf:Person”) of “foaf:Agent” and the sub-classes (i.e., “prov:Organization”, “foaf:Person” and “prov:SoftwareAgent”) of “prov:Agent” (Fig. 6).

---

Fig. 6. Re-using existing ontologies for the GCIS ontology. Each ontology is marked with its unique color and namespace prefix.

To improve the interoperability, use of existing ontologies in GCIS ontology leverages the existing datasets based on them (cf. Simperl, 2009). By adopting and adapting existing ontologies and/or mapping classes and properties of GCIS ontology into existing ones, such as those asserted by the “rdfs:subClassOf” property in Fig. 6, the resulting GCIS ontology is compatible with existing ontologies. This is of great value because the dataset (i.e., triples in RDF format) based on the GCIS ontology will also be interoperable with other linked data resources (e.g., DBpedia\textsuperscript{24}, GeoNames\textsuperscript{25}, NPG Linked Data\textsuperscript{26}, etc.) in the Semantic Web, and thus the provenance mining with multi-source datasets will be facilitated. The role of provenance mining is to find hidden provenance information of agents, entities and/or activities in the “Web of Data” (cf. Simmhan et al., 2005). For example, a viewer sees the metadata description of the dataset used in a paper, then by clicking one or more buttons/links on the user interface, the GCIS can retrieve and present more information about this paper from the NPG Linked Data, and information about the instrument used to generate the dataset from DBpedia.

3 The GCIS ontology (version 1.1)

A primary purpose of the iterative ontology engineering approach was to permit humans to obtain a clear understanding of the concepts under consideration, and for computer applications to encode the semantics in a form of syntax (i.e. define classes and properties and set up a framework to organize them). In the use case analyses, we transferred new classes and properties to a GCIS ontology concept map, which was used a “white board” for domain experts and computer specialists to discuss issues such as naming and defining new classes or properties, reusing components or structures from other ontologies, mapping classes and properties into other ontologies, etc. We then encoded that concept map into RDF. In this way, the GCIS ontology has been refined by version updating (e.g., we currently have a version 1.1) along with the use case-driven iterative approach, and we keep the RDF document and the concept map synchronized.

3.1 Basic classes and properties

The primary classes and properties of the GCIS ontology (version 1.1) are shown in Figs. 7 and 8. Fig. 7a was partly derived from the result of Use Case 1. Comparing with Fig. 2b, the classes and properties in Fig. 7a are extensively enriched with inputs from other use cases. Another significant difference is that in Fig. 7a we asserted classes “gcis:Image”, “gcis:Dataset” and “gcis:Organization” compared to the use of “dctype:Image”, “gcis:Dataset” and “org:Organization” in Fig. 2b. We asserted those classes in order to avoid the so called “ontology hijacking” behavior (Hogan et al., 2010), in which third-parties modify external classes/properties that have already been defined. For example, in the GCIS ontology (version 1.1) we asserted “gcis:Image” as a sub-class of “prov:Entity” (Fig. 8a). If we replace “gcis:Image” there with “dctype:Image” we will assert “dctype:Image” as a sub-class of “prov:Entity”, which is regarded as an inappropriate behavior of ontology hijacking.

Fig. 7. Basic classes and properties in the current GCIS ontology (version 1.1): (a) Classes and properties representing a brief structure of Draft NCA3; (b) Classes and properties related to the findings of Draft NCA3 and each chapter in it; and (c) Classes and properties about sensors, instruments, platforms, and algorithms, etc. that datasets are derived from. The figure does not cover all the classes and properties in the GCIS ontology (version 1.1). For a full version of the concept map, readers can refer to: http://cmapspublic3.ihmc.us/rid=1LZJTQWPM-1R2VB58-2LVG/GCIS_Ontology.cmap.

We defined “gcis:Finding” and associated classes and properties (Fig. 7b) because the final NCA3 will contain a list of findings at the report level and each chapter will also present findings at the chapter level. Classes and properties displayed in Fig. 7c were partly derived from Use Case 3 (Fig. 5). We also referred to the Semantic Sensor Network (SSN) Ontology\(^{28}\) in building out this part of the GCIS ontology (version 1.1). The current RDF (serialized in Turtle (Terse

RDF Triple Language\textsuperscript{29} format) of the ontology is accessible on our website\textsuperscript{30}, together with a document describing it (see footnote 27).

\textsuperscript{29} http://www.w3.org/TR/turtle/. Accessed on 2013-09-25.
Fig. 8. Mapping classes and properties in GCIS ontology (version 1.1) into those of other ontologies: (a) A few classes are asserted as sub-classes of “prov:Entity” and “prov:Activity”, respectively; (b) A few roles in Draft NCA3 are asserted as instances of “prov:Role”; and (c) A few properties are asserted as sub-properties of “prov:wasDerivedFrom”, “prov:wasAttributedTo” and “prov:actedOnBehalfOf”, respectively. This figure does not cover all the mappings. For a full version of the concept map, readers can refer to:

We also identified more roles and asserted them as instances of “prov:Role” (Fig. 8b) after the analysis of Use Case 2 (Fig. 4). Figs. 8a and b are high-level conceptualizations of the provenance graph described in Fig. 4a. Besides mapping classes and instances into other ontologies, a few properties in the GCIS ontology are asserted as sub-properties of corresponding properties in the PROV-O ontology, such as “gcis:hasChapter” is asserted as a sub-property of “prov:wasDerivedFrom”, “gcis:hasAuthor” as a sub-property of “prov:wasAttributedTo”, and “gcis:subOrganizationOf” as a sub-property of both “prov:actedOnBehalfOf” and “org:subOrganizationOf” (Fig. 8c). A number of datatype properties such as “rdfs:label”, “rdfs:comment” and “dcterms:identifier”, etc. are omitted in Figs. 7 and 8 due to the space constraints. Readers may refer to the RDF file and the document explaining the ontology (see...
footnotes 27 and 30) for further details. We will also put updates about the GCIS ontology engineering work online\footnote{http://tw.rpi.edu/web/project/geis-imsap/GCISOntology. Accessed on 2014-02-07.}.

### 3.2 Deploying the GCIS ontology in pilot systems

We are deploying the GCIS ontology (version 1.1) in several pilot systems. Our aim in those systems is to present Draft NCA3 both through a human-readable web site as well as a machine-readable interface for automated mining of the provenance graph. We introduce two of the systems here. The first is a Draft NCA3 linked data browser (Fig. 9)\footnote{http://globalchange.tw.rpi.edu/standalone/geis/report/nca3draft. Accessed on 2013-09-26.}. To build this system, we first transformed Draft NCA3 from PDF files into structured TXT files and then into RDF triples (Fig. 10)\footnote{For technical details, see: http://tw.rpi.edu/web/project/geis-imsap/NcaTxt2Rdf. Accessed on 2013-07-31.}. We then set up a triple store into which we loaded the RDF triples\footnote{For technical details, see: http://logd.tw.rpi.edu/tutorial/installing_using_virtuoso_sparql_endpoint. Accessed on 2013-07-31.}. Finally we used the Epimorphics Linked Data API Implementation (ELDA)\footnote{https://code.google.com/p/elda. Accessed on 2013-07-31.} to build templates for querying the triples and a user interface for browsing the results.
Fig. 9. Homepage of the pilot system “Draft NCA3 linked data browser” (as it appeared on 2013-09-26).
Fig. 10. RDF triples describing Draft NCA3. Classes and properties defined in the GCIS ontology (version 1.1) are used in those instance triples.

The homepage (as it appeared on 2013-09-26) of the first pilot system is depicted in Fig. 9, which shows the page of Draft NCA3 and chapters in it. Clicking a chapter identifier will open a page of that chapter containing information on the chapter title, author roles and chapter findings, etc. Then, clicking an author will open a page with information about that author, such as author name, affiliation, and contributions to Draft NCA3, etc. In this way, the system traces the origins of each component in Draft NCA3 and presents them to viewers. By using the GCIS ontology in the first pilot system, each instance has its resource type (e.g., report, chapter, and figure), identifier and persistent landing webpage, and the links among instances are established. Those links also become a part of the description of an instance on its landing webpage.
Fig. 11. A list of figures in Draft NCA3 with information about the chapters they are in and the images they have. The pilot system was developed with the GCIS API (version 0.61).
Fig. 12. Landing page of Chapter 2 in Draft NCA3. The pilot system was developed with the GCIS API (version 0.61).

The second pilot system (Fig. 11)\(^{36}\) referred to the design of the GCIS ontology (version 1.1) for content representation and organization. It used both a relational database and a SPARQL endpoint\(^{37}\) as the back end. Fig. 11 shows the retrieved information about all the figures in Draft NCA3. The result can be in JSON, YAML or HTML format. Fig. 12 shows the landing page of Chapter 2 in Draft NCA3, from there the metadata of that chapter can be output in several formats. Through using the system a viewer can conduct provenance tracing such as which figures are in a chapter, which images are included in a figure, the creation date for an image, the


spatial and temporal extents of that image, the location from which to download the physical file of that image, and the datasets used to generate that image, etc. The SPARQL endpoint supports provenance tracing queries. The next section lists a few examples.

### 3.3 SPARQL query examples

We are collecting data from Draft NCA3 according to the classes and properties defined in the GCIS ontology, and loading them into the SPARQL endpoint of the second pilot system. Knowing the ontology used in the endpoint, users can draft novel SPARQL queries for provenance tracing. Below are two SPARQL query examples with the currently loaded triples.

Draft NCA3 has a number of chapters, and each chapter has a number of findings. The first SPARQL query example is to find all the findings of Chapter 2 in Draft NCA3.

```
PREFIX gcis: <http://data.globalchange.gov/gcis.owl#>
WHERE {
  ?c a gcis:Chapter ;
    gcis:chapterNumber "2" ;
    gcis:hasFinding ?f
}
```

As described in Use Case 3, a figure in Draft NCA3 may consist of a number of images, and in turn those images can be derived from a number of resources such as publications and datasets. Publication has various types, such as articles, reports and books, etc. The second SPARQL query example is to find all the articles in the resources from which Figure 2.26 in the Draft NCA3 was derived.

```
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX gcis: <http://data.globalchange.gov/gcis.owl#>
```
4 Discussion

Recent studies (Fox and Hendler, 2009; Villa et al., 2009) show that there is a gap between domain scientists (e.g., earth and environmental scientists) and Semantic Web researchers. The former are growing ever more dependent on the Web for retrieving resources, collaborating with colleagues and publishing research outputs, but have no coherent agenda for exploring the emerging trends and potentials of the Semantic Web technologies. On the other hand, the latter have largely focused on formal aspects of semantic representation languages and general application developments, with inadequate consideration of requirements from specific science areas. The use case-driven iterative approach we applied to this work has proven useful to unite different parties and to apply the latest Semantic Web technologies to topics of interest to domain sciences. We hope this method can be applied further in the field of Earth and environmental sciences.

The use case-driven iterative method has an iterative procedure, and it advocates ontology re-use and supports rapid prototypes. Those features are similar to a few existing methods, such as the PrIMe (Miles et al., 2011), the Content Ontology Design Patterns (Gangemi and Presutti, 2009) and the Agile Methods (Cohen et al., 2004). However, the use case-driven iterative method has its unique features that make it especially suitable for developing domain-specific Semantic Web applications. The method highlights knowledge engineering rather than software engineering,
although the latter is also covered in its scope. It includes domain scientists in the working team rather than isolating them as product users. The PrIMe (Miles et al., 2011) is a guided approach for provenance-aware software engineering. It requires the developers to consult the application users to understand what kind of provenance information is requested and at what level of details. The use case-driven iterative method, in contrast, addresses the collaboration of a group of people with different roles, such that the end users (e.g., domain experts or stakeholders) are a part of the working team. This ensures each step in the method is understandable and acceptable to end users, and makes it easier to develop and implement domain-specific applications.

Content Ontology Design Patterns are small ontologies that mediate between use cases and ontology design solutions (Gangemi and Presutti, 2009). By implementing that method, a resulting domain ontology is a composition of a few content patterns, plus appropriate links and extension. The work of re-using and mapping to existing ontologies in this paper is similar to the method of Content Ontology Design Patterns. However, the ontologies re-used in our work are not ready-to-use small ontologies (i.e., building blocks). We first had to review the structure and pattern of them and then adapted a part of the classes and properties into the GCIS ontology. The Agile Methods (Cohen et al., 2004) face software engineering, so a team implementing Agile Methods consists of software developers. The use case-driven iterative method, in contrast, faces knowledge engineering for Semantic Web applications and it advocates a working team of different roles. Because the Agile Methods face software engineering, it encourage quick and frequent changes where appropriate (Cohen et al., 2004). We applied the use case-driven iterative method for ontology engineering, so we assembled knowledge components from use case analyses and made the resulting ontology stable. The difference also reflects in works of documentation. The Agile Methods put documentation as the last option as long as the communication within the group is effective. For the use case-driven iterative method the documentation of each use case analysis is essential, because they are necessary for tracing how each part of the resulting ontology is derived.

As mentioned above, we referred to existing best-practices in the leveraging and extending of ontologies to improve the interoperability of the GCIS ontology and its capability for provenance representation. A primary existing ontology we re-used in this work is the PROV-O ontology.
We also investigated the Open Provenance Model (OPM) (Moreau et al., 2011)\(^{38}\) and the Proof Markup Language (PML) (Pinheiro da Silva et al., 2006)\(^{39}\). In our previous works we had applied OPM (e.g., Tilmes et al., 2010) and PML (e.g., McGuinness et al., 2009) for representing provenance of Earth and environmental science data. We re-used PROV-O in the GCIS ontology for three primary reasons. The first is that the PROV-O is more comprehensive compared to OPM and PML. In Table 2 we extended the Michaelis et al. (2009) comparison of OPM and PML and added comparison to classes and properties from PROV-O and ISO 19115. The table shows the similarities between those models and ontologies, and it also shows the improvements from OPM and PML to PROV-O, such as the definition of property “prov:wasAttributedTo” for relationships between an entity and an agent. The second reason is that, as a W3C recommendation, the PROV-O represents the consensus from a broader community. Mapping the GCIS ontology into PROV-O will improve the interoperability of our ontology and the datasets underpinned by it in the “Web of Data”. The third reason is that the use of PROV-O is currently rare in the community of Earth and environmental sciences. We want to conduct further investigation of provenance capturing by using the PROV-O ontology and collaborate with the PROV-ES group\(^{40}\) to work on a specific provenance ontology for Earth and environmental sciences.

---

