http://goo.gl/jRZLF

Krishna's Presentation on Use Case ***As a PDF, images preserved***

Link to video presentations (pending)

NSF Meeting - Face - to - Face Workshop

Drawing the roadmap for the semantic/ontology based infrastructure for Geosciences
April 30, 2012

1. Purpose: Introduction, including community(ies) to be served, technical area(s) of the roadmap, and brief discussion what improvements in the present state-of-the-art in geoscience data discovery, management, access, or utilization it will enable. Also include examples of how the outcomes from your effort will enable the community to be more productive and capable. (Panelists: Peter Fox, Boyan Brodaric, Naijun Zhou, Calvin Barnes)

APS: Serve the needs of scientists and citizen; technical areas: model/ontology, knowledge, data, services, tools to integrate and analyze
NZ: Data sets can be used differently, not just semantics (knowledge should be looked at as well); community based approach-agreement within a community of the use of data and semantics. Can be a community of the communities, and everyone can join the community (as a user or researcher).. Need a vision for Earth Cube in terms of semantics and ontology.
Q; What are the communities to be served? Atmos, ocean, earth sciences, geography (bordering communities such as biology, ecology)

Discussion Item for website: Discussion around what the composition of the community is... who is involved not involved.. ask NSF's guidance

Search and Discovery -> More complex applications; Roadmap can have an evolutionary path

Use cases are not always known beforehand, use cases must be powerful enough to get the needed buy in. May need to capture sufficient context of the data for it to be utilized by other communities.

Data discovery can be a trigger to participation:
to put more data that is relevant to the the research that earth-scientists are performing into their hands easily, without them knowing anything about the underlying semantics resulting better research. to have them find data they never knew was there.

What will be the incentive to contribute data? Limited data sharing, lineage of the data (attribution), data as a product is important.

Here is a paper about a survey that addresses incentives and obstacles to sharing data
http://www.sciencedirect.com/science/article/pii/S1574954112000222

As a data technician, how can I name a new variable to work with the ontologies? Information should be interoperable with the largest community, tags must be provided to allow for this . You don't need to change your schema but map to the tags (schemata) provided by others.

Use case discussion: Use cases limit our vision; there are other approaches to use cases; carefully composed, cross-cutting use cases (and use-case templates) can drive development of applications which may initially appear to be more narrow but eventually lead to needs expressed in the larger vision, thus they can be quite powerful, USGS mineral set - if you can describe mineral using USGS mineral vocabulary. data search use case is easier with OWL but changes in other cases (e.g., integrity constraints checking needs modifications to the formal semantics), use case templates can help make the whole structure more manageable

Quality Assessment: Semantics is the key to quality assessment of data - enabling the ability to do quality assessment and to propagate the knowledge needed across brokering and throughout layers is an important component of EarthCube capabilities

Timeline: roadmap is useful to develop realistic timeline (low hanging fruit -> longer term projects), if earthcube is designed using use case then cases must be carefully defined, open environment alternative, seems to be consensus on short term goal of greater data availability for scientists, long term goal knowledge availability, are there metrics to characterize success based on timeline? quantifying success

Data Conservancy has some life sciences/earth sciences use cases developed, if EarthCube would like to see them.

When talking about purpose, needs must be looked at;

Q:How will this group define the scope?
A: Scope of the roadmap is semantics, community based on EarthCube vision
Terminology needs to be approached

Q:How does this roadmap fit in with other roadmaps?
A:No harmonizations of roadmaps before June event/keep in mind relationship to other groups/what forms the framework?

Anne: It seemes to me that there are two roles: the data provider and the data consumer, and three communities: research, citizen scientist and policy...

2. Communication: Description of a communications plan with end users, developers, and sponsors, as well as links to and feedback from other EarthCube community groups and EarthCube concept projects to promote systems integration and accelerate development. Include a discussion of needed interactions with allied fields, agencies, and other related activities (present and desired).(Panelists: Anne Thessen, Krishna Sinha, Tim Finin, ?)

NOTES
Have a dedicated person to deal with communication (Facebook, blog, Twitter, website); keep thumb on community (right people get invited to meetings workshops).

In terms of workshops, there is an upcoming workshop on a very related topic, Workshop on GIScience in the Big Data Age, held in conjunction with the Seventh International Conference on Geographic Information Science 2012

Make sure that people know that they are welcome to participate at many levels, not just at the "ontology" level. Some work at controlled vocabulary, glossary, taxonomy levels. Lots of good, important, useful data available at all these levels.

Repository of vocabularies and ontologies, enables people to get information about what groups are involved

Who are involved with other EarthCube groups? Outcome of meeting - list of liaisons to other EarthCube groups

It will be important to engage research domain community through demonstration of results; e.g. documenting success stories. converting non-believers and communicate the power of the technology (TED talk)

Managing a negative image to non-believers, lesson learned - "expectation management"

Obvious things to do include developing a common Web site for our effort to evolve a semantic infrastructure for the Earthcube community, creating a multi-author blog for news and comment and establishing a set of public mailing lists (with archives). Links to tutorial papers, assignments

Vocabulary camps that bring ontology engineers and domain scientist together to really develop ontologies, document, and publish them; see http://vocamp.org/wiki/Main_Page . See http://vocamp.org/wiki/GeoVoCampSB2012 for a geo specific camp as example. There is an upcoming GeoVoCamp in Dayton: http://vocamp.org/wiki/GeoVoCampDayton2012

Community efforts like MMI (Marine Metadata Interoperability) or SOCoP (SOCoP.org) very important. marinemetadata.org EarthCube should promote the development of communities of practice within domains and also have "community of practice" teams to promote ways to harmonize interdisciplinary semantic interoperability. And S&O community of practice should be represented in the EarthCube governance.

Challenge: EarthCube site is RDF A enabled, (an open standard for embedding structured data directly in HTML. This allow posts to be automatically tagged using FOAF and Dublin Core) can take SPARQL queries

Solicit ideas for use cases on the group's website, then use cases can be searched

Communicate more directly to what the community needs to know/misunderstandings. Presentation: "*Engaging the public in earth sciences: Five guiding principles*" by Edward Maibach, M.P.H., Ph.D. (GMU) http://wiki.esipfed.org/images/3/38/Maibach_ESIP_Jan_'12.pdf One point was "Use audience research to determine what to say." "What to say is determined by

the needs of your audience, not by what you are most eager to say."

MPS: we have the ning site which does a lot of the above, but it has been a bit hard for me to find things quickly on ning-- too many nooks/crannies. E.g. couldn't easily find the semantics discussion wrap-ups from the Charrettte...

Possible outreach activities are to organize one or more workshops in conjunction with appropriate conferences (e.g., X, Y and Z), propose special issues on semantics for the geosciences for relevant journals, invite someone to write a short (one or two page) high level article to be submitted to an appropriate venue and ensure there is some representation or connection with relevant standards organizations and their subcommittees (e.g.).
education outreach

Invite bioinformatics gps for lessons learned

A structured Body of Knowledge(BoK) helps education and communication. See GIS&T Body of Knowledge as an example.

TEXT

Communication within a large, interdisciplinary project such as EarthCube is extremely important and difficult. Time and care will be required to facilitate communication between domains, such as geosciences and computer sciences. An infrastructure needs to be in place to facilitate communication within EarthCube and with users and stakeholders outside EarthCube. A good communications plan will include a position dedicated to disseminating information. Activities include writing press releases, managing website content, writing and publishing a newsletter, managing audience expectations and maintaining a social media presence. Within and between project communication can be achieved through dedicated conference lines, Skype, instant messaging and face-to-face meetings. It is recommended that project wide "all hands" meetings happen annually, face-to-face. At these meetings, projects can give status updates and provide feedback.

More targeted communication and/or education opportunities can be exploited by holding workshops during major scientific meetings or helping to sponsor a scientific meeting. User surveys can capture the audience needs/wants to help further target our message and goals. We recommend the development of "Communities of Practice" to keep the lines of communication open wide between EarthCube and domain scientists. These CoPs would

3. Challenges: Description of major drivers, trends, and shifts impacting or that could impact the focus of a working group, including but not limited to changing technology,

adoption culture, and community engagement. (Panel: Ruth Duerr, Pedro Szekely, Xiang Li, Cyndy Chandler)

Diversity: There are a lot of communities and diversity, diversity in terminology between communities. There are also diversity of scales, formats, levels of data (metadata vs data). Diversity on all scales for EarthCube in generally

Adoption: Semantics -pay first - get benefits later, why should researchers learn new things and mark up their data?

Development of Ontology: especially when having to develop ontologies for broad areas.

Trends: rapid increase in types of new sensors, is a challenge, but could work in our favor

Multiple perspectives, citizen science, sensors, heterogeneity, multi-modality and multiple levels of abstraction. People's perception

OWL technology challenge - "human" data capture, non specific of definitive data capture,

A formal language (like OWL) cannot capture everything we need. Semantics is not (just?) about OWL (or any other specific technology), many other modeling approaches play a role as well (e.g., human language, mathematics, abstract algorithms, executable software systems, etc.) However, the semantic web gives us all the great tool support.

Heterogeneity should be seen as something to be exploited not resolved

We don't have the tools we need to do this. We need to evolve and mature these tools. A problem is that too much time, training, and effort is currently needed to define semantics well. Better front end tools and ontology visualization may help with this. We should expect to pilot the the use of such tools in the first 3 years of the roadmap. An issue is what effort to put into developing/enhancing these tools, such as Protege, as part of the effort and how to select candidates.

Building ontologies is hard, boring, and tedious --- go after the low hanging fruit, use work that has already been done. (added later: but then we miss the opportunity to fully develop the capabilities that s&o can provide, so we should make sure the ultimate system is defined in timeline)

Drivers: science is increasingly interdisciplinary, need to find ways to effectively communicate results to broader community showing the value and benefits of "semantics", bioinformatics as a driver, benefits recognized when science is shown differently, acute need to deal with real world problems, deluge of data, models, and tools to use (overwhelming, what's the best choice?), geoscience work has impact on climate, hazards, decision making, open api's, open data, cannot rely on a small number of computer scientists to develop for you

lessons learned from bioinformatics can help in dealing with challenges

People don't want to have to reinvent the wheel, interoperability for models and tools used, people can register their resource against the ontology

Do we agree that semantic web and LOD is a driver here? Life sciences has a big presence here, Do we mention LOD in the roadmap? Suggested that bio-informatics is "throwing it's hat in the ring of LOD, but note that lots of bio ontologies that are not useful for LOD.

Development of LOD API's will be a significant driver in this community, in the same way that adoption of OWL as a standard back in 2004 changed things. We are now talking about real problems with real data.

Acts of Congress, national mandates, there has been one being worked on for the past 10 years

4. Requirements: Process(es) to be used to get the necessary technical, conceptual, and/or community (i.e., end-user) requirements at the outset and during the life of the activity, including approaches to achieving community/end-user consensus. (Panel: Mark Schildhauer, Xiang Li, Pedro Szekely)

<MPS>: article talks about identifying processes to get necessary technical, conceptual, and community (End-user) requirements, but practically we should reverse order, first talk to domain community to assess needs, from this develop the concepts, and lastly identify the relevant, enabling technologies.

There are a lot of impressive projects currently out there, which EarthCube should try to capitalize on. To develop Requirements, first Need Thematic workshops involving domain scientists facilitated by semantics experts. Some degree of cross-cutting generality might be criterion for a suitable Use case, since EarthCube is intended to be more generalized than a highly specific use case. Then, we need Semantics workshops involving KR engineers, etc. to compare/communicate about various technology approaches and their strength and limitations. Development and promotion of ontology design patterns may help enhance interoperability of semantics approaches within earthcube. Earthcube should also develop or promote a standard mechanisms for annotations. Identify best of class, highly useable tools to assist in semantically annotating data. Also, “advertise” foundational frameworks among EarthCube participants, that might be useful to standardize upon. Finally, identify incentives to keep participants enthusiastic and involved.

There is need for active involvement from other groups on EarthCube, and related efforts such as DataONE, NEON, Data Conservancy, NCAR, OOI, IOOS, etc.</MPS>

One of the processes to be established is to have the community communicate use cases. Where are they feeling pain? Also throw in challenge problems? Direct funding to these problems. Challenge problem (could be an agreed upon use case)

A challenge problem can be used to measure progress along the way incrementally.

One approach would be for the semantics working group members to develop conventions for documenting both domain and infrastructure use-cases and then work with other EarthCube working groups and the larger communities to “encode” their use-cases in a consistent manner. Such activity would present opportunities to embed concept maps and other semantically interesting components. Requirements could then be derived from these.

Workflow group has developed a long survey. What would a semantic survey look like?

Who is the survey being sent out to?
We need to take into account knowledge infrastructures. Semantic enablement layer of

knowledge infrastructures also needs to happen, e.g. the ability to exchange information between OGC Web services (Spatial Data Infrastructures) and the semantic Web.

Collecting use cases can be a hopelessly difficult exercise. Use model catalog instead

Two different types of use-cases: domain use-cases, abstract general infrastructure use case such as how search works.

General EarthCube challenges: How can I find the data that is relevant for my purpose? discovery, assessment, access

5. Status: Description of the state of the art within the topical area of your roadmap. This should include approaches and technologies from geoscience, cyberinfrastructure, and other fields, the public or commercial sector, etc. that have the potential to benefit the EarthCube enterprise. (Panel: Anne Thessen, Peter Fox, Boyan Brodaric, Clinton Smyth)

Life Sciences: names, names management, digitization, mining, modeling species morphology semantically. On verge of retirement scientists will be willing to submit data via brain dump at MBL.

The major challenges for the life sciences revolve around liberation of data from text and the long tail of small providers. The state-of-the art for the former is TaxonFinder and Neti Neti for finding taxon names and CharaParser for extracting morphology information. The latter will prove to be much more difficult as it requires buy in from the community. Annotators and data repositories exist, such as Dryad (which works with publishers), to try and capture long tail data. We look to GenBank as an excellent example of a data repository that has excellent community support. The state-of-the-art in life sciences semantics includes projects like TaxonConcept, which represents species on LoD and Phenoscape, which, as its goal, performs inference to find evolutionarily important genes.

Life Sciences are unique in that taxonomic names are an important and near universal piece of metadata. The state-of-the-art in names management is represented by the Global Names Project and the Encyclopedia of Life demonstrates the beginnings of what can be accomplished by managing data around names.

Next steps for life sciences include, but are not limited to, semantic species descriptions, automated generation of reconciliation groups, semantically-enabled identification keys and text mining for information about species interactions.

Geology: development of specific formats, ontologies in different field, semantic similarity measures & matching, mapping within linked data. Challenges: large volumes of remotely sensed imagery Most of the data sits on individual PC's. Data is lost after someone leaves/retires/hard drive fails. Vast majority of data creation is individual scientists. How can that data be shared/integrated/discovered worldwide? What is the . response from the engineering community? Discoveries are made by the individual scientist. Getting into the shared data space is the challenge. Raster data has challenges when publishing data.

Tools: need for tools that make this interesting to the domain scientists, need an editor (not available), small projects but more accessibility is needed

Computational sciences: Advances in Web scale information retrieval, big data management, cloud computing, massively parallel computing, information extraction from text, machine learning, and semantic web technologies are enabling many disciplines, including the geosciences, to obtain and analyze more data. Do these belong in other special interests group? Some of these changes in the computational area enables interesting things with large amounts of data.

Emergence of workflows to be semantically driven.

How do machines come into play? A lot of the semantics are in the terms. What is the need for heavyweight semantics that is also machine understandable? Is this in the focus of the project?

Scalability is forefront in Life Sciences. Natural language processing and machine learning algorithms. How do we scale to the millions?

Semantically driving machine processes for experiments, instructing the robot to redo experiment based off of interpretation of results. New ideas can be generated

Current ontologies were produced by humans process knowledge to create semantic structure/nodes. Getting the knowledge in, in a semantically consistent way is a huge bottleneck.

Status should be addressing things from some of the other sections. Tuning technology to capture doctor's bedside notes into semantic

Controlled English, restricted english - domain experts can express input that then

becomes knowledge formalized by logic in the system.

CMaps - cognitive maps, informal link and node mechanism; Pat Hayes Univ. of W FL, OWL converter

Ontology Extraction- going beyond language extraction, goes into relations and events, looks for entities and descriptions and how they relate taxonomically , sometimes can be easier to obtain than arbitrary events

Curation is required. Workflow with a review process where people are responsible for curating the data and the semantics that describe the data.

Overview of e-science and the idea of the executable paper? Documenting workflows so they can be re-run? Process ontologies, etc. Oppys here , there are some general ontologies for science knowledge/ reasoning techniques to look for new ideas, the rest is probably domain specific

Evaluation requirements, a matrix may be beneficial. For example, some applications are real-time that may need “dynamic” ontology, thus a sucessful solution using predefined ontology may not work well here.

Scalability - TB, PB of data, studies with user interaction show that slower application response tend to be less likely used. Bigger data systems (Facebook, Twitter) system must be scaled horizontally rather than vertically, this can handle increasing user demands. vertical scaling is not a cost effective long term solution. tweaking and simplify semantic models and inference rules as a way increase performance is not cost effective solution and may be considered a form of vertical scaling. Horizontal scaling may including sharding and distributing queries and rules engine across a distributed triple store.

Approaches: The informatics approach involving small teams of domain experts, data curators, and computer science professionals working collaboratively to develop the infrastructure that meets the requirements identified through use case development is ‘state of the art’.

Linked Open Data vs. Data in DB tables - research being done on how to enable DB to appear as linked open data

25% of LOD is geospatial (geographic?) data (or at least contains geographic reference,

e.g., places or locations)

Use of cloud is now widely used, hadoop etc. better ability to handle bigger data, number of semantic tools that are available is growing, free services available, substantial progress in infrastructure

Informal, individual scientist data needs long term stewardship.

Who wins? How is the young scientist incentivized?

The VLDB Conference committee assigns experimental reproducibility labels to accepted papers called reproducible and shareable.

- Reproducible label: the experiments reproduced by the committee support the central results reported in the paper.
- Sharable label: the experiments are made available to the community and they have been tested by the committee - a URL is provided.

In some biomedical communities, data and results must be submitted. Results added o community knowledge store

How can we change to status to what we want from the future? European community encouraging data submission different from the text. Data will go straight into a database.

Data sharing is a controversial topic (in part since there are many alternative ways that try to address it).

Geoscience information presentations

Status on metadata- we need to raise the bar. Currently, the minimum bar is of the type required in FGDC. With the capabilities enabled by web services infused with S&O, we can raise the bar to include information sufficient to assess data quality.

6. Solutions: Process for the identification and comparison (pros and cons) of approaches and technology solutions that will contribute to the EarthCube goal of satisfying current and future research needs of the geoscience end-user.(Panelists: Amit Sheth, Pedro Szekely, Ruth Duerr)

======= SUMMARY =====

The process of identifying and developing solutions call for realizing that we have two communities to serve – the community of users, and the community of developers. The community of users will generally not be interested, nor should be interested in specific technologies (eg semantic Web technologies) underlying their applications and solutions. They will likely be exposed, however, to the ontologies that describe the concepts they will use in posting queries or questions to a system that will help them get information and insights they need. Even if the users are themselves not interested in details of semantic technologies, it would be valuable for them to be aware of "semantics-empowered" (analogy is "Intel inside") message, so that gradually they are a participant in making the solutions more powerful by being better users of semantics and by providing better semantics in the form of domain knowledge that can make systems incrementally more powerful (e.g., through ontology evolution).

The developer community should be aware of and whenever possible use relevant standards and community developed/adopted specifications that have good tooling support, such as those from W3C's Semantic Web initiative and other community efforts in Semantic Web (RDF/RDFS, OWL, SPARQL, Linked Open Data-LOD), W3C's Semantic Sensor Networking, (Semantic)Web Services (SAWSDL, SA-REST, WADL), OGC's Sensor Web Enablement, and ontologies such as SWEET. It should also learn from and adopt with appropriate changes architecture of successful systems in other domains, that encompass ontology development and evolution, semantic annotation for a broad variety of data (text, images, video, sensor data of various modalities, etc) and models, semantic search/browsing/filtering/querying and advanced semantic processing and reasoning (path and pattern finding, inferencing) in support of complex analysis, discovery, problem solving and decision making.

EarthCube community should support development and public sharing of open source vocabularies and ontologies, semantically annotated data, workflows and tools, use cases and best practices, challenges to evaluate diverse solutions to a common problem, and demonstration of successful applications and systems serving end user (scientist) needs. The latter can play a key role in adoption. Support for a semantic infrastructure or resource for EarthCube community that can host a registry and provide common services, such as one patterned after National Center for Biomedical Ontologies for the biomedical domain, can be considered. Quality of data and annotations, including support for provenance should be a key capability of such a resource. LOD has already become a key semantic approach for data sharing, and is anticipated to play a major role in EarthCube community too.

===== NOTES TAKE AT THE WORKSHOP =====

One approach would be to look for earth science solutions that use semantics, perhaps in a narrow domain and see if they can be expanded to other domains.

Competition is the traditional mechanisms for deriving solutions, is deciding what the solution is and then building effective? If the solution cannot be adopted by all, by definition solution cannot be deemed successful. However, the IETF and W3C are models of successful community efforts in developing large-scale solutions via promoting standards. E.g. witness success of RDF and OWL (W3C recs)

Hosting data in LOD could provide more access for the data to be used by others

Having an announced competition could give a winner, semantic web community already does challenges-- but in this case, could tailor competitions to practical needs of EarthCube.

Success can be measured in linked data on how often they are used. Adoption measurements can be a convenient way to view success

Two audiences: the semantics wg has two audiences, developers and science users - earth scientists. It is important that we speak openly about semantics within the developer community as solutions are developed. However, a large majority of earth scientists have little interest in semantics (and there is no reason why they should). For these earth science community scientists, semantics needs to be delivered “under the covers”. However, with useful tools for these users, that have semantics included, it is important that we advertise that they have “semantics inside” so they begin to see its value.

Semantic technologies that get infused are the ones that really address the science drivers.

Explicit demonstration of semantic value in applications, can bolster adoption

Earth science community has less risk than biomedical community, scientists buy-in to semantic uses, semantic web is not new anymore

Anne: Tools developed in life sciences could be translated to rock/mineral naming, How essential are the names to data management in geology? Cal: Instrumentation ontologies have/are developed. Large amount of data sets are already stored/sequestered on individual computers. Rocks and minerals are the fundamentals of geology, these need to be addressed in a semantic way to get community adoption

Semantic ecosystem would consist of ontologies/vocabularies, annotation tools and publication of annotations, tools to integrate and analyze and visualize. Credit can be given to other data sets that are used with your own when describing results

Vocabularies, annotation tools, organizing data - where do we need solutions in these individual parts?

Driven by current needs and future needs. Current need - get the data out there, how can we make it easy for them to share data, Future - Where will the Geoscience community be in the future? Data without borders, what solutions that need to be developed can make this happen?

Another current need that semantics can help with is the ability to discover “data of a known quality”. My experience is that when scientist’s in my organization become aware of data related to their field of high enough quality they will get that data “by hook or crook” - semantics or not.

“Design and Development of Linked Data from The National Map” (USGS) online at: http://www.semantic-web-journal.net/content/design-and-development-linked-data-national-map

Semantic side is one of the weakest sides of the USGS case

Earth sciences semantic is going to be similar to biomedical sciences semantics, leadership about the importance of semantics

NSF best approach - bring community together, to ask what the best solution is, convergence around ideas and governance set up for development process, is this possible in a much larger geosciences community?

An approach from data.gov a way to start? encourage pdf online -> csv, sql dump -> add semantic information. Moving from 0-star data to 1-star is a bigger jump than 1-star -> 5-star data.

We would like to capture non-funded data, a lot of data is not in numerical spreadsheets, but field data, descriptions, pictures, sketches

Barbara NSF - Challenge: Come up with brief information that you can post on website for geoscientists, explain semantics, importance in lay terms that anyone can understand, more end users will start to be directed to EarthCube site

Biomedical community has set up community registry for ontologies, loose governance structure, community guidelines on how to build ontology... thought there would be more discussion on community registration to get further along

This process from biomedical community has been supported by millions of dollars from NIH. The process costs money.

Proposal - Separate groups that are already building semantically enabled applications (continue working on these) but there is a need for external communication and coordination about the uses of semantics within these. A Geek Squad could be established by EarthCube to go between these groups. Initially they would gather information about the technologies are being used, how effective they are, and how will the applications work for the users. In later instantiations of the geek squad it would take on other roles such as educating different groups about what “best practices” are being used in the larger earth-cube community of semantic developers.

Have people with real data/real problems to reserve slot with Geek Squad to actively engage scientists to solve problems, scientists do not need to understand all of the details of knowledge representation language but they need to understand what we model in the ontology. A ‘Geek Squad’ could organize ontology workshops to engineer ontologies together with domain scientists based on their data and needs. The Geek Squad also needs to actively engage the community.

Open Ontology Effort - see http://socop.oor.net/ for the NSF funded geospatial ontology repository (built on the BioPortal software) which could also store Earth, Geo and GI Science ontologies. This could be an interim approach and is ready to be used now. The SOCoP OOR is is one of several OOR instances that are designed to be federated. We might create separate GIScience repositories and populate these.

http://mmisw.org/orr/#b hosted by Marine Metadata Interoperability (MMI) Project https://marinemetadata.org/

in bioinformatics- semantics added to web portal with shared resources

Look at some successes in biomedical and understand the differences in geosciences, data is not as accessible in bio, we need to start with data and build semantics from data? Look at successful investments (PubMed, etc)

Geosciences differences - more fragmented community, uncoordinated duplicated efforts, need more coordination, there are some shared ontologies,

Lessons learned from the Semantic Web working groups within the ESIP Federation and NASA Earth Science Data Systems Working Group: Having the user community come to the Geek Squad (this does not happen) users are not going to seek semantic solutions. lesson learned - need to continually engage the user community, user community - scientists, and IT people | Geek Squad continues through EarthCube dev. Another lesson learned includes making the user community a part of the working groups to maximize the effectiveness of bridging the gaps between the technologist and science communities.

Travel budget for IT tech people, as they need to be in these rooms as well

Geoscientists being a soft science - semantics is the thing that is the hope of the planet, looking for patterns

What research questions require what types of ontology engineering?
Light/heavyweight ontologies

What role can the agencies play in developing this? What do you need help from NSF /other agencies to achieve these goals?

-Some do not like this idea, EarthCube must come up with guidelines for semantic registrations

-Cliff Jacobs - Thought Experiment - You really need to be able to suggest a process, hopeful that you are able to come to some ideas on this, how that is implemented can be a variety of techniques that can be submitted in proposals

-Data availability - semantics can be part of the solution but it cannot be the whole solution. The easiest way to make scientists make data available - NSF/publishers require this

-LOD/semantics make data availability useful for future use

Geospatial data is one thin layer of data that we would like to see with semantic registration. Scale issues need to be addressed (nm -> km) information needs to be developed at many different scales

-- registry for (searchable, discoverable) data, ontologies/taxonomies/vocabularies and factual/background knowledge; how far the data will be discoverable, reusable?
-- support for data generated by device/sensor observation as well as citizen sensing (social networks, citizen science)
-- how will it support collaboration to support science involving more complex problem solving (because of common vocabulary, semantic interoperability, more power analytical tools enabled by use of semantics and integration) compared to what predates EarthCube
--
Part of the solution is to come up with tools and communities of practice that encourage capturing semantics at the source where the best knowledge of how the observation came to be committed.

7. Process: Process(es) to develop community standards, protocols, test data, use cases, etc. that are necessary to mature the functionality of the topical area and promote interoperability and integration between elements of EarthCube. (Panelists: Calvin Barnes, Cyndy Chandler, Naicong Li, Philip Murphy)

How do we organize ourselves to make progress? What can be done? What's our role?

Short term - come up with solid roadmap for June meeting

Mid term - communication,

Long term - How far do we want to go here?

Feasible, sustainable, deliverable

Clarify the issue of scope - geoscience? earth science? what things? data, models, tools

Addressing data is not enough, addressing process is important. Concern about buy-in from diverse community, get past "s" word, introduce semantics in a straight forward way, large groups already at AGU, [ESIP Fed[1]], GSA, ways to use 1 or more use cases

[1] http://wiki.esipfed.org/index.php/Semantic_Web

at national/international meetings to address issues dealing with semantics

An application that demo's semantics gets scientists interested in semantics

Develop a use case out of this group

Janet: there are 2 communities: those who want to get the data out there, well-described to "feed the Cube", and another community who want to address the discoverability and linking communities of data ..
How do we develop a process that makes it easy to be a 'good citizen' wanting to share data, - put money into tools, what are the underlying technologies that can be useful, how do I adopt these software, infrastructure.

Is there an active datanet project for geosciences? OCI, We should review those activities to look for similar focus areas. Strong synergy possibilities-- or partnering of EarthCube with DataNet program to develop cyberinfrastructure and semantics in support of GeoSci. Is this a possibility?

Semantics is not necessarily embraced by entire science community. Forming a working group that focuses on semantic infusion would help to bridge the science community with the technologist communities. Infusion processes includes assessing what are the science drivers, current capabilities, pain points, and wishlists?

One charge - develop a use case scenario that bridges disciplines , needs to be fully transportable to other EarthCube working groups, intellectual liaison

Work with early adopters to develop use case, engages members of research community, (DataOne, ESIP)

What are the characteristics you are looking for from the use case?

Highlight the need for recognizing heterogeneity , improve our thinking about models

Like the idea of canvassing the community for use cases, first slide highlights some principles for what would be a good use case

Most of the vocabularies we are considering are terminologies. This is lightweight semantics and a good starting point. The slides recently shown demonstrate the idea of reasoning and using semantics to discover new knowledge and make implicit facts

explicit. This allows you to have a formal axiomatization, reasoning and deep semantics.

Krzysztof: Biodiversity is an interesting example as it cannot be modeled in OWL. It needs a logical pattern to do so. Mark (Krzysztof's friend) asserts that biodiversity community is indeed working on an OWL implementation of Darwin Core, the main "biodiversity" metadata standard. So-- the semantics of what is even meant by "Biodiversity" is at question here...Krzysztof: which is a great example showing how different people arrive at a different understanding and use different methods. I am looking forward to see the work done on this in Darwin Core. MPS: agree wit ur pt. Anne: I'm wondering exactly what is meant by not being able to represent biodiversity on semantic web. Do you mean that biodiversity metrics cannot be represented? Krzysztof: sorry for the confusion, my point was much simpler and not about biodiversity as a research area but just a (over)simplified definition of the term, e.g., something along the lines 'The variety of life (the diversity in terms of different species) in an ecosystem'. For us, it is an interesting observation that such a definition cannot be directly modeled in OWL. Anne: okay, i see, thanks.

Can not use taxonomies that can not be resolved in OWL.

Starting to get a consensus that this might be a good use case. Question is what is the process to develop consensus about a use-case within the community.

Krishna's Presentation (as a PDF images preserved) on a suggested usecase on *volcanism and climate*.

One approach for earthcube might be to create a registry that could be use to store use-cases (possibly through the ESIP Federation). We probably want to drill down in semantic use-cases to more detail. Broad spectrum of these use cases. Discovery use cases are important, if they can find new data relevant to their research, they are immediately interested/excited.

NSF survey is a good source of information for use cases (200+) Most are currently open ended and difficult to organize. Use case templates can help with this. Convert use cases to templates is not easy. Template should have parameters related to semantics, tie this to other working groups efforts.

If you identify use cases, how does that contribute to moving the entire enterprise

forward? - When you dev use cases that provide direction to areas of content development, which ontologies need to be developed for which domains? It indicates what need to be developed, how to use the ontology to organize data, models,

The use case is the manifestation of the process, way to engage everyone

Use cases:

1. one or two sentences describing a use case (as in Krishna's slides)
2. one or more paragraphs in a more narrative format (as expressed in the many Expressions of Interest contributed earlier)
3. use-case templates - a short form with several components, including: summary statement, step-by-step description of the normal flow of the system being developed, an activity diagram as a pictorial representation of the steps , concept map that would eventually develop into an information model. The words used in the use-case and the concept map set the stage for semantic analysis and then development. Not all of use-case types 1 and 2 would have to be expressed in templates.

The use case integrates with semantics in two ways: how the semantics is required and how the semantics developed then can be contributed to the broader community.

Use case can help solve immediate problem, communicate need for semantics to scientists; each use case semantics will overlap and work together

Cliff Jacobs: What we are looking for from this group, understand the process to move something forward to sustaining a working group. 1. Have to have broad community engagement, if you want to build infrastructure 2. Risk mitigation - rather than send us your best ideas probably will not move to earthcube goal, spend money now to mitigate risks

Having a repository is a good idea and important, next section is timeline

8. Timeline: Timeline for the project and all related sub-projects, including prioritization of activities and measurable milestones/major achievements and total resources (human and financial) required to achieve roadmap goals over a period of the next three to five years.

Short term: By Next Week (through June 15th 2012)

- Identify new temporary committees and members online
- Every panel that met here, through internal communications, come up with a document
- All ten documents available for review by everyone
- May 7th Conference Call 2-3pm EDT
- May 17th - webex - collab document is available
- May 30th - document goes to NSF
- May 7th Identify the people who can make connections to other EC working groups.

Year one
-
Year two
-
Year three
-
Year four
-
Year five
-
Year ten
-

9. Management: Management/governance/coordination plan and decision-making processes necessary to successfully establish standing committee(s) and subcommittees (if warranted), including a plan to identify and respond to shifts in technologies and changing needs at the end-point of use. Include discussion of approaches to educating end-users and achieving community consensus on advancing the capability/technological solution.

1. Collectively author manifesto on website for S/O group -explaining why S/O is important and backing the explanation up with two or three use cases.
2. Establish use case registry
3. Establish project registry
4. Develop matrices - comparison matrix of projects from registry, compare/contrast existing semantics technologies, ontologies vocabularies
5. Problem with current Earthcube Ning site - NSF possibly move to opensource community framework? S/O should endorse this move.

6. Engage communities like ESIP, EU, Australia

1. Need to establish some committees - can move faster on topics
    a. Outreach - engage the semantics and its challenges, deliver our vision and message to other communities (eg ESIP/federal agencies)
    b. Technology - monitoring status of semantic technologies - what's coming up; what is a sufficient definition of this? Has someone already described this (ESIP)? There is an existing semantic technology roadmap generated by the ESDSWG.
    c. Joint Committee (Domain & IT People) - use use-cases to identify targets of opportunity with goal of delivering low hanging fruit to community; similar to the Infusion committee (within ESIP and ESDSWG); part of this work is identifying pain points, make the solutions count where the infused semantic solutions directly address gaps in the science communities.

Concern - Why would I want to be a member of new semantic web community? Better to join up with existing communities?

Committees can be formed closely to the matrices previously proposed

How do the committees fit into the whole process?
http://wiki.esipfed.org/index.php/Semantic_Web
10. Risks: Identification of risks and additional challenges to the successful establishment of any working group, and any unique risks associated with a working group associated with your topical area. With respect to identified risks, an approach to risk mitigation should be addressed. (Krysztof Janowicz, Naicong Li, Philip Murphy, Isabel Cruz)

1. Risk or benefit to hide semantics from domain experts?
2. Ontology engineering- lightweight/heavyweight, under vs over engineering, knowledge engineering bottleneck
    a. one risk is not utilizing all of the available capabilities
    b. incompatible expressivity-- SKOS vs OWL-DL.
3. Community not adopting - solution strategy to prevent community from not adopting
    a. usability testing - done early and often but often takes a backseat
    b. perception that ontology is not correct (created by non-domain expert)
4. What if the technologies change over time? make systems robust
5. Communicating
6. Semantics serving other groups? Semantic awards could be reduced as other groups mention semantics too
7. Semantic web tech people don't understand domain user's needs
8. Risk in Letting the opinions of domain scientists drive the technology solutions-- (e.g.

would domain scientists have endorsed adopting relational databases or HTTP/HTML?). Best informed computer science/engineering should drive technology, along with knowledgable domain practitioners whose needs must be well represented in Use Cases.
    a. can start with big data franchises - NOAA, USGS, NASA, building semantic solutions, and then these capabilities can inform engaging individuals and giving them useable tools (e.g. how Mosaic catalyzed Web access) semantically annotating their individual data sets (those millions of spreadsheets prevalent in Geology, etc.)
    b. more difficult task is to deliver tools to individuals
    c. But-- simultaneously informing domain researchers about how semantics works is important. Risk is not moving forward until critical mass of support and understanding from domain rschers.
9. Ontology redundancy - petrologists can define rock the best rather than other earth scientist
    a. an ontology is not always right or wrong, but more or less useful?
    b. Information loss in translation from natural language to formal language

10. How often does the ontology need to be updated? Ontology evolution?

If done right, users will not know about semantics in the backend system. But some aspects of semantics are exposed to users such as semantic annotations. Need to continue to reduce barrier of entry and outreach to science community. Need to ease users into being more comfortable with semantics.

Retaining current people (trained students) is a challenge

What are some of the metrics we could use (short-term, long-term) to measure success?
- look at linked open data statistics, algorithm that shows how well you fit in with LOD

Semantics (RDF) has been identified as huge cost savings for DOD way to reduce risks/costs (references: semantics for data sharing and semantics for interoperability)
-reduces risk of evolving systems
-greater agility in engrg to reduce risk

Mitigate risks though agile sw dev, if fear is scientists are afraid of semantics, consequence of failure is non-adoption; involve community of domain scientists
-rocks in medical ontologies - two approaches to address interoperability

Risk--- opportunity cost of NOT doing semantics
Risk -- we miss the opportunity to utilize the full capabilities of S&O for quality assessment, as well as discovery and linking data

Risk mitigation: understand the risks that other semantics communities faced (e.g., the biomedical community).

Definition of use cases is a huge part of the roadmap

- Failure to engage the broader geoscience community
- Developing overly complicated ontologies that are difficult to understand and use
- Failure to track de facto standards for semantic representation as they evolve
- Ignoring important kinds of data (e.g., images, signals, videos)

References

One of many overviews on current status of Semantic tech: Semantics Scales Up

Webex #2

E-mail sent by EarthCube discussed - announcement on EarthCube website

Volunteer for Group sub-committees

Joint Geoscience/Tech Committee - charged with developing use cases, identify selected use cases

- Hassan Babaie
- Calvin Barnes
- Anne Thessen
- Mark
-

http://goo.gl/JVRLM
Committees formed today need to have results by Wednesday