EarthCube Design Approach

Aaron Braeckel (braeckel@ucar.edu), Bob Barron (bob@ucar.edu), Arnaud Dumont (dumont@ucar.edu), Bruce Carmichael (brucec@ucar.edu)

National Center for Atmospheric Research (NCAR) - Research Applications Laboratory (RAL)

26 September 2011

Introduction

Several related but independent data infrastructures are under development, including the NOAA 4-D Weather Data Cube and EarthCube. These systems share a common theme: distributed, universal, discoverable access to information. Despite the breadth of federal agencies, detailed requirements, and data types being addressed, there is a set of common requirements and usage patterns. One of the core unification points between domains is that of geospatial and temporal information. Although fundamentally different in nature, geological and climate data can be jointly understood through the lens of their shared geographic characteristics.

The NOAA 4-D Wx Data Cube and the FAA NextGen Network-Enabled Weather (NNEW) systems are independent, distributed systems that will interoperate using shared data access and data distribution standards. The authors of this whitepaper utilize the experience in developing and working with this NOAA/FAA system as the basis of the concepts described in this paper. It is noted that the requirements for the EarthCube go beyond those addressed in the 4-D Wx Data Cube and NNEW, and therefore the approach has been extended to address the educational, scientific and research requirements of EarthCube.

Concepts

Users - Members of the EarthCube community which may include the general public, educators, researchers, students, and scientists from across the geo-sciences. Users may also include groups from other related systems, such as the 4-D Weather Data Cube.

Community of Interest (COI) – A group of users who share a common interest within EarthCube. Examples include scientific disciplines such as atmospheric science, oceanic modeling, or aerosol measurement), technical domains (Fortran developers, supercomputing, grid computing management), organizations (NCAR, University of Washington), or other types of groups. COIs are related/linked to other COIs, such as the atmospheric science community being related to the atmospheric chemistry community. Each COI has a group of users who are responsible for managing the COI. COIs track lists of unsolved/critical problems, document how to solve common problems, and provide reviews of results. A user may be a member of multiple COIs.

Resource – A resource that is made available to users of EarthCube. Resources may be considered the nouns of the system, the *things* that are acted upon and shared. Resources come in many forms, such as: data files, algorithms, research results, visualizations, proposals, educational materials, white papers, documentation on research opportunities, best practices, and reports. Resources may be made available to either all or a portion of the EarthCube community. This allows researchers to expose early results to a limited group, for example. Resources may be related to other resources, such as a dataset used in a publication.

Service – a service made available for the use of the EarthCube community. Services may be considered the verbs of the system, the *actions* that are taken on resources. This includes making resources accessible, processing and computing capabilities, mass storage, performing resource transformations, and disseminating data. Services come in a variety of forms from machine-readable, remotely-accessible web services (SOAP, REST, etc.) to manual human-oriented services (such as emailing staff for access to mass storage, a physical phone call to arrange processing tasks, etc.). Web services are hosted in a distributed fashion by members of EarthCube. All services may have quality, usage and other metadata associated with them.

Registry – a web service that provides access to metadata, such as discovery metadata and data heritage metadata. A Registry is roughly synonymous with a Catalog. Due to the fundamentally-distributed nature of EarthCube, the top-level Registry must support federated capabilities (queries distributed across the system).

Vision

The EarthCube is envisioned as a system where "cyberinfrastructure enables the geosciences". The EarthCube will allow users to be conductors, using a palette of resources, processes, and communication to compose their work, reducing time spent resolving tedious problems. Facilitating data storage, processing, retrieval, and transformation are all key elements of the system-to-be. Common, time-consuming processes like re-projection and file format conversion are well-understood problems that can be (and have been) solved by CI and should not obstruct science, research, and education. Common problems do not have to be solved repeatedly.

EarthCube will also be a collaboration hub, allowing users to discover and share resources, information, and contacts. This allows users to announce their resources, interests, and work to other users. EarthCube will support and enable interactions between the geosciences, from terminology and units of measure to data format and metadata.

Much as the Internet itself is diversified and distributed across multiple organizations and individuals, EarthCube must rely on and enable distributed resource management and access. Due to existing infrastructure and management realities, it is neither possible nor desirable to centralize the broad variety of resources used in the system.

No system will be heavily used unless it is convenient to use, regardless of the power and flexibility. Self-describing capabilities, the Unix-like composition of simple components into more powerful processes, and web-based interfaces all often lend themselves to a system that is easy to understand and convenient to use.

"The Resources in the EarthCube are Discovered and Accessed so that they may be Collaboratively linked with other Resources in convenient ways that enables new Contributions, all in a Trusted environment."

Discovery of resources, services, and collaborators

One of the critical aspects of any broadly distributed system is the ability to track and enable discovery of resources, services, and collaborators. On the open web much of this need is satisfied with search engines, which are a critical component to everyday Internet activities. Users and systems must be able to discover resources, services, and collaborators of interest across the entire distributed system. For public resources, discovery must be enabled through search engines on the open web; text search does not need to be re-implemented. The system must also remain aware of what resources, services, and collaborators are available.

Access to resources and services

Due to investments in existing infrastructure and the variety of both communities and requirements, it is not envisioned that a single set of access protocols will meet all needs. Rather, a smaller set of core, capable, geospatially-enabled, standard protocols will be encouraged alongside the full diversity of existing protocols. All access is not equivalent, and HTTP and FTP is not enough when efficient access to resources is a priority.

Collaboration of users, processes, and resources inside and outside of the geosciences

COIs embrace different management policies, data formats, terminology, units of measure, and metadata that must be bridged for EarthCube to facilitate data sharing and higher-order capabilities. Collaboration also includes communication of proposals/researchs, knowledge sharing, process improvements, and notifications of updates made by users or communities.

Contributions from the EarthCube community

EarthCube must anticipate and allow for the evolving needs of its community. The user community of EarthCube itself is seen as the most important source of innovation and the most valuable self-organizing resource. Contributions come in the form of data, reviews and quality information, documentation, data services, resource transformation capabilities, etc. EarthCube shall act as a platform to enable, facilitate, and govern contributions from all parts of the EarthCube community. It is envisioned that a set of components of EarthCube will need to be identified that should be pluggable, rather than every aspect of the system being highly mutable. EarthCube itself must remain stable while supporting change.

Trust in the system as a whole

Users of EarthCube must be able to trust in the utility, quality, and adaptability of the system to their uses. An unstable, unreliable or untrusted system will not be used, no matter how functional. EarthCube must address data integrity, reliability, access control, and data provenance as critical elements of a usable system. Users must trust that if EarthCube does not meet current needs, that it will evolve to satisfy their requirements.

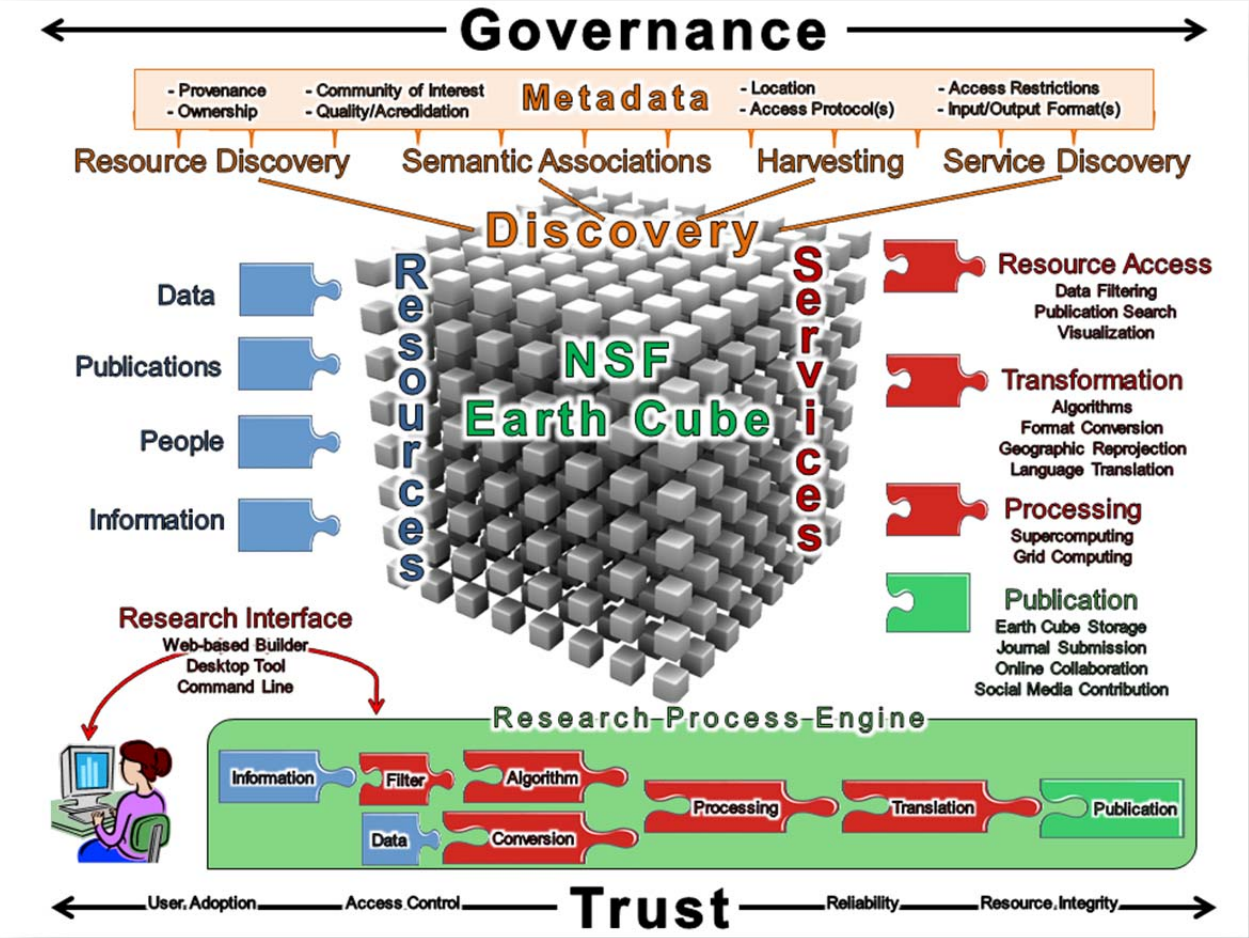

Figure 1 - EarthCube vision

Capabilities

To fully realize the vision above, the following implementation-independent capabilities must be supported:

COI and User Capabilities

- Notifications – users and COIs shall be able to register their interest in changes to resources from other users and COIs. For example, a user wants to be notified whenever a particular resource is updated and would also like to be notified whenever new publications are posted to the meteorological modeling COI.
- Resource requests – users and COIs shall be able to note their interest in resources that are not currently available. For example, a COI needs a high-resolution terrain model, a user needs a Python script that can transform GRIB2 data files, or an organization would like to request research/papers on a particular topic
- Home page – users and COIs shall be able to create a home page that lists resources (such as data and papers), relationships to other users/COIs, allow for communication between users, and provide a place to document and discuss topics of interest (wiki, blog, etc.)
    - Publishing of:
        - Papers, reports, and ongoing research results, and the data associated with them

        - Proposals and calls for submissions
        - Collaboration opportunities, including available expertise and current research areas
- COI forums – COIs shall be able to host or link to community forums, email lists, and other mechanisms for community support to enable question/answer interactions

Resource Capabilities

- Plotting/visualization – map- (geo-) based quick look with political boundaries and other existing datasets, plotting templates (provide input file and axis names, for example), generation of a tag cloud image from a document, etc.
- Transformation – temporal re-sampling such as unifying/correcting obs data times and sampling input data to a fixed timestamp (01:00Z), spatial resampling, coordinate reference system/projection transform, geometric transform (translate, scale, rotate), perspective transform, transpose, shear, warp, resample (nearest, bilinear, trilinear, bicubic), data quantization (restrict to a fixed set of values), data type transform (int32 to double32, int32 to byte) with quality options, units of measure conversion, etc. Transformation capabilities are realized in several forms: web services, command-line calls (documentation of existing libraries), and links to downloadable libraries
- Data type conversion – grid to contours, wind U+V to wind speed and direction, tabular (CSV) data to vector or grid data, regularly spaced grid to ragged arrays, matrix to N-dimensional grid, etc.
- Format conversion – NetCDF to CSV, Word to PDF
- Language translation – English to French, etc. Especially useful for papers and other primarily human-readable documents
- Usage metadata – how often and who has used particular resources. This is particularly useful in providing information about what resources should be retained and whether time and money should be invested in expanding or maintaining them
- Quality metadata – data quality, precision
- Other metadata – provenance, responsible user/organization, history of transformations
- Semantic markup and associations – information contained, relationship to other resources
- Quality control checking – units-based verification (negative precipitation, relative humidity outside the range 0-100, negative wind speed, incompleteness, accuracy, precision, missing/unknown, etc.

Service Capabilities

- Distributed data access
    - Support for request/response and publish subscribe message exchange patterns
    - Real-time data feeds as well as static, fixed datasets
    - Access to resources with protocols amenable to devices such as mobile phones
    - Standard access for reliability and security (HTTP/FTP)
    - Next generation access for improved efficiency and capability (geospatial services such as WCS, WFS, WMS)
    - Support for a variety of protocols, such as BitTorrent for large file transfer
- Federated registry and metadata access – COIs may choose to manage metadata relevant to its domain and host its own registry that participates in the EarthCube federation
- Access to offered services such as supercomputing services, grid computing services, and mass storage
- A Unix-like command line tool for composing web services that is usable both on the web and locally. This tool is itself a service hosted by EarthCube, so there may be more than one interface on the capability and the utility may be used by automation. Here “resample”, “nc2shapefile”, “subscribeWFS” and “grid2contour” represent a command that interacts with a web service:

earthcube:~> resample http://foo.org/myfile.nc -crs “EPSG:4326” > ftp://bar.org/results/myfile-4326.nc
earthcube:~> nc2shapefile ftp://bar.org/results/myfile-4326.nc > grid2contour > /tmp/myfile.shp

earthcube:~> subscribeWFS http://foo.org/services/pubsub-wfs -f myFilter.xml -o ~/fooWFSData/
earthcube:~> cd http://foo.org/services/pubsub-wfs; ls -l #lists capabilities and offered resources

Discovery Capabilities

Discovery of:

- Resources: data files, imagery, papers, publications, etc.
- Educational materials
- Services: geospatial data access services, transformation services, security services, HTTP/FTP services, etc.
    - Computing Resources (supercomputing, grid computing, etc.)
    - Storage Resources (mass storage)
- Transformation/Processing/Evaluation capabilities
- Users
- COIs
- Semantic associations/markup on services and resources

Security and Integrity Capabilities

- Integrity – data and/or messages cannot be modified without detection
- Confidentiality – unnecessary information is not divulged to third parties
- Availability – services and information are available when needed
- Identity – users and COIs may be identified

Governance

It is recognized that there are many communities that must be able to internally manage their resources and services, and these capabilities must also be fit within the greater EarthCube community. One of the greatest challenges will be to take advantage of the existing capabilities/investments of the disparate communities and enable commonality and collaboration.

A COI is the best authority to manage the semantics (ontologies, for example) of its community. These semantics are related to other COI semantics, but are fundamentally managed in the COI.

COIs will be able to mark up services, resources and collaborations with classification metadata. For example, a user might be identified as an administrator of a COI by classification under EarthCube/COI/administrator. A dataset provided by a particular service may be marked as the authoritative source for certain uses of the COI by classification under EarthCube/COI/authsource/usecase. All classifications are managed by the relevant user or COI.

Existing resources and services shall be integrated in an inclusive manner, but contributors will be encouraged to move to more commonly useful mechanisms. For example, HTTP and FTP are a form of data services that are well understood but do not allow for efficient geo-spatial queries. The least common denominator among the geosciences is geospatial and temporal information, so while HTTP/FTP data access will be supported contributors will be encouraged to support more capable and efficient geospatial services. Similarly, all data formats shall be welcomed in EarthCube but contributors will be encouraged to make data available in commonly used, metadata-capable data formats. Encouragement shall take the form of advertisement of the capabilities of a

resource/service during the discovery phase. For example, it should be apparent to users looking for data services whether transformations would have to deal with the original data volume, or that a particular resource format is not commonly used and would need conversion.

It is not yet clear to the authors whether it is practical for EarthCube to integrate with existing user and group identity and authorization management systems. It is not always possible to federate identity, trust, and authorization across a community as broad as EarthCube and therefore new EarthCube-specific capabilities may be required.

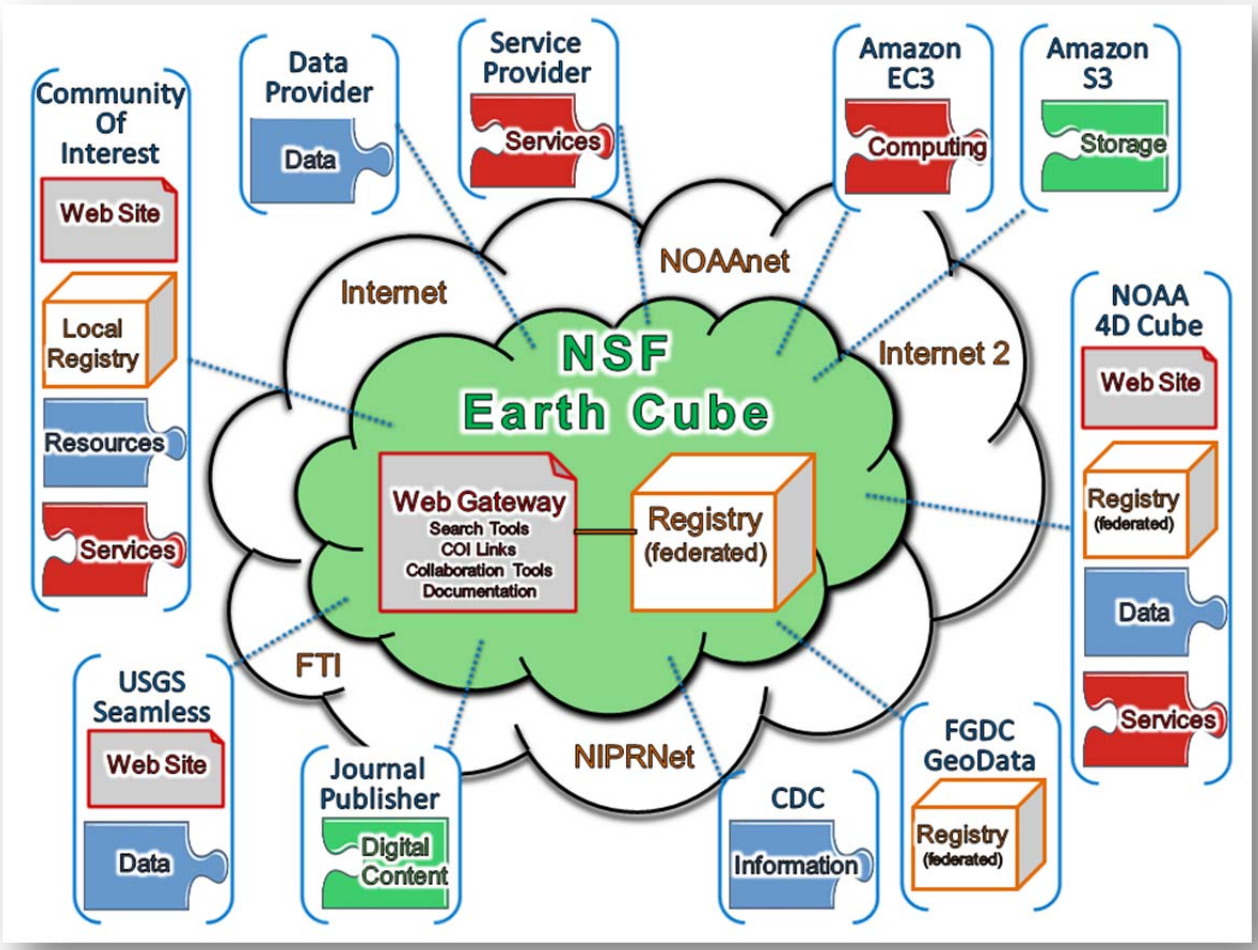

Figure 2 - federated content

Architecture

This is an implementation- and technology- independent starting point for components of an architecture that satisfies the capabilities described above. A specific architecture and solution would realize the following functional components as a web site, Java application, or other specific libraries and applications. The components below are not necessarily an exhaustive inventory of what might be needed to enable the capabilities described above.

- Master registry and sub-registries - a master EarthCube Registry would federate queries to sub-Registries. The master registry is an NSF cloud-hosted service, and sub-registries may be hosted by COIs, users or organizations
- Monitoring software – software components that periodically collect service availability, quality of service and similar metadata that may impact the operation of the running system. This process watches the running system and actively prevents systemic issues. Service quality metadata is placed in the corresponding registry. System maintainers may also be notified of critical issues
- Metadata harvesters – software components that periodically contact services to collect usage, quality, and resource metadata. This is used to gather and keep metadata about the system in sync with the system itself. Harvested metadata is placed in the corresponding registry
- Data Access Services: HTTP/FTP, geospatial data services (WMS, WCS, WFS), BitTorrent, etc.

Resource Transformation Services as described in the

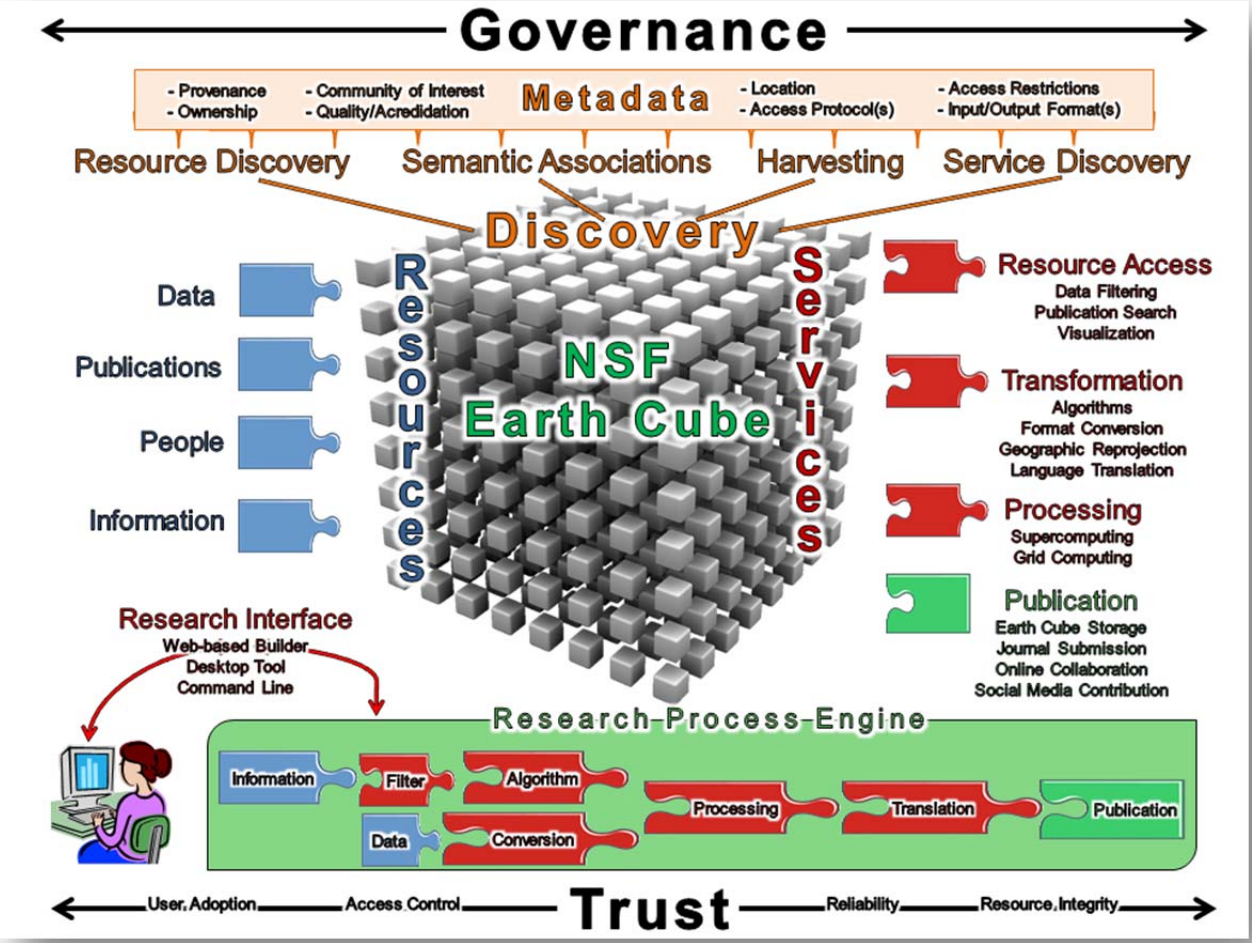

Figure 1 - EarthCube vision

- Capabilities section such as:
  - Units of measure conversion
  - Data format conversion
  - Human language translation

- Etc.

- EarthCube Gateway (http://earthcube.gov). This is a web site hosted in the cloud by NSF that provides default mechanisms for discovery, COI management, and access to all the capabilities of EarthCube. Other web (or non-web) gateways/portals could also be supported in the long run that address specific use cases, each of which could make use of EarthCube registries and services

Implementing default mechanisms for the user-facing capabilities described in the

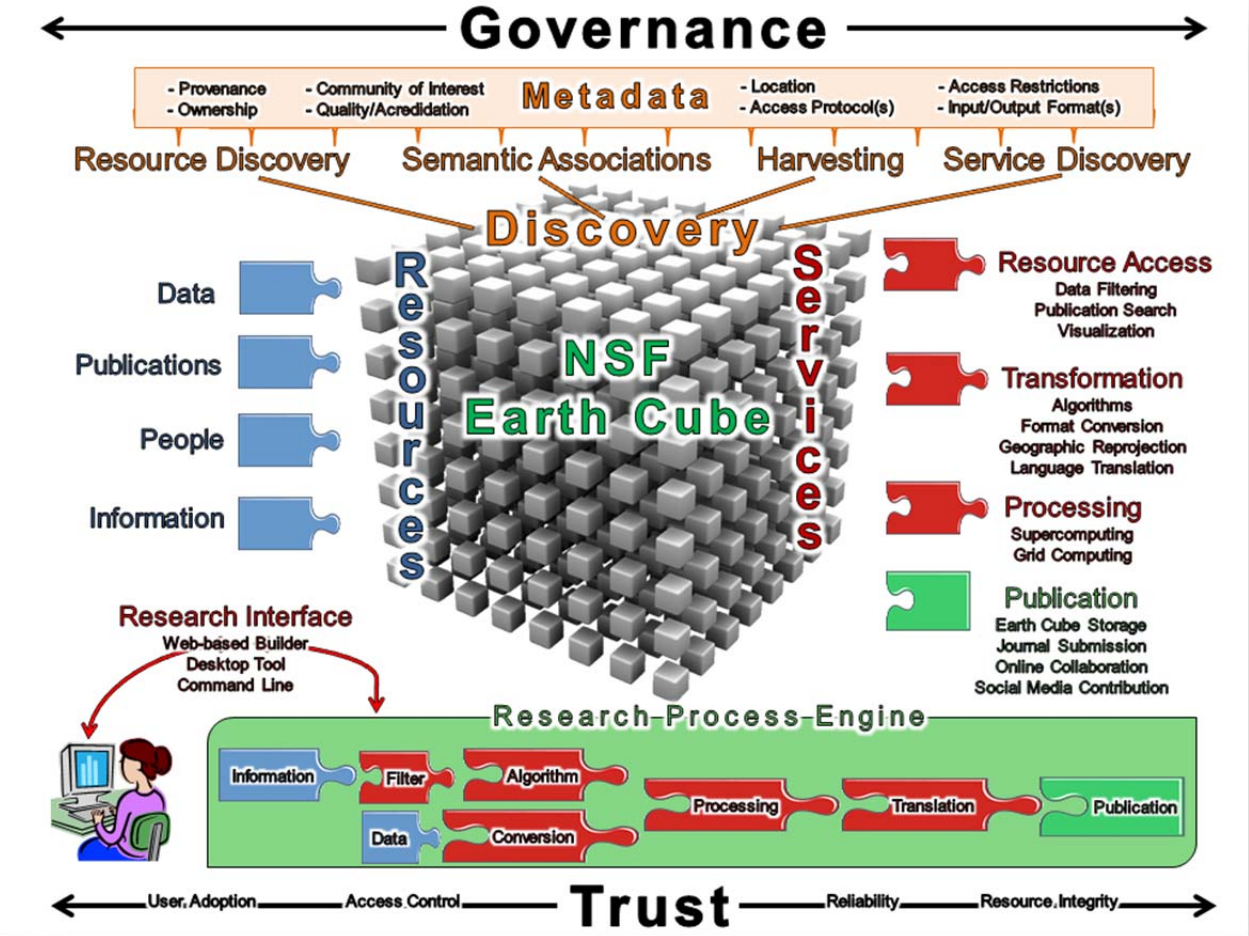

Figure 1 - EarthCube vision

- Capabilities section such as:
  - An interface for user registration
  - Documentation on EarthCube itself
  - Discovery of users, COIs, resources, and services
  - Simple resource viewers
  - Semantic Viewer and Editor – sub-communities may manage their own semantics, and these may be linked to the semantics of other sub-communities
  - Resource publication mechanisms – email, "What's New" newsletter, Atom feed, etc.
  - Etc.

- User/COI identity management system
- Access management system – a service that tracks and provides access to information on what users/COIs are allowed access to what portions of the system
- Units of measure system
    - UoM formal definitions (parseable by automated systems)
    - UoM identification metadata for different resource formats

Design Process

Each step of the following process shall have a defined time period and shall allow members of the community to provide feedback and input. Some activities may be performed concurrently.

1. collate and synthesize "raw" use cases into a consistent and well understood set
2. distill and prioritize requirements based on the use cases - what takes priority: cross-domain data access, visualization, or open access to supercomputing?
3. document core use cases into an EarthCube CONOPS (concept of operations) document. Experience on the FAA NNEW and NOAA 4-D Wx Data Cube has shown this to be an extremely useful and heavily utilized artifact
4. evaluate candidate solutions against use cases
5. develop a mature governance model. This must predate integration of existing system solutions
6. compile a list of criterion for candidate system solutions. For example, candidate systems must be 100% open source, must be operational for over 6 months, must be cross-platform (Windows, Linux), etc.
7. gather a list of existing systems that address one or more of the architectural elements above – registries/catalogs, data services, transformational capabilities, visualization, etc.
8. Evaluate the ability of these systems to interoperate in a cohesive EarthCube
9. identify gaps that require new implementation work
10. solicit and award proposals from the community that address specific gaps
11. repeat:
    - solicit requirements, issues, and feedback on the current EarthCube system
    - re-prioritize requirements
    - identify gaps
    - solicit and award proposals

Team membership has not yet been finalized. Final determination is pending developments and further information at the EarthCube Charrette such as the specific skillsets of attendees. NCAR/RAL has experience in the aviation, climate, and meteorological domains. NCAR/RAL has years of system design and development experience on the FAA NNEW program and NOAA/NWS 4-D Weather Data Cube, including many areas in distributed systems including OGC geospatial data access services, web services more generally, service-oriented architecture, metadata, registries, data formats, data visualization, and standardization activities.

Information on the NNEW program and the 4-D Wx Data Cube may be found on the NNEW wiki[1]. This wiki contains a wealth of documentation (design, CONOPS, requirements, white papers, etc.), technical discussions, and software.

[1] http://wiki.ucar.edu/display/NNEWD

Operations and Sustainability

As EarthCube is primarily envisioned to be a layer above existing capabilities, therefore it is important that EarthCube automatically take account of changes to those capabilities. Metadata harvesting is a key example of this need; there are many systems whose metadata is not able to keep up with changes to the available data and capabilities of the actual system. EarthCube should accurately understand the current contents and capabilities of the system itself. It is the experience of the authors that asking contributors to manually keep metadata up to date with the running system is not reasonable or practical.

With federated capabilities comes the responsibility to ensure that the capabilities that are broken, missing, or unreliable are addressed. Monitoring will be used for gathering availability information, usage information, and system policy enforcement.

Cloud computing is a nebulous term and is being used to describe a variety of capabilities such as outsourced email services, hardware hosting, virtualization, and full software development platforms. As described in the Architecture section, it is envisioned that core EarthCube components be hosted on cloud infrastructure as IAAS (Infrastructure As A Service), in this case hardware hosting and OS virtualization. Identity and access information will need to be evaluated carefully for its suitability on a public cloud due to security and confidentiality requirements, it is possible that a private or government community cloud will be necessary. Hosting and virtualization techniques increase the potential scalability of EarthCube, and allows for higher operational reliability. To support uptime and scalability requirements, the critical central components of the system such as the registry must be developed to support high availability and load balancing.