EARTHCUBE CONCEPTUAL DESIGN
A Scalable Community Driven Architecture

http://earthcube.org/group/scalable-community-driven-architecture

Overview

PI: G. Djorgovski (Caltech)

CO-I: D. Pilone, T. Pilone (Element 84), D. Crichton, E. Law (JPL)

Other key personnel: S. Caltagirone (E84), S. Hughes (JPL),

T. Huang (JPL), A. Mahabal (Caltech)

2016 ESIP Winter Meeting
A high level system blueprint for the definition, construction, and deployment of both existing and new components to ensure that they can be unified and integrated into an evolutionary national infrastructure for EarthCube
Methodology

- Identification of stakeholders, concerns and requirements
- Identification of architectural use cases and drivers
- Selection of an architectural framework
- Development of the architectural principles
- Development of the architectural models
- Capture of the architecture artifacts in a consolidated report
- Generation of recommendations for adopting the architecture for the EarthCube program
## Stakeholders

<table>
<thead>
<tr>
<th>Stakeholder/Actor</th>
<th>Concerns</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSF Program Managers</td>
<td>Make decision and provide guidance at the EarthCube program level.</td>
</tr>
<tr>
<td></td>
<td>Provide sufficient funding to support the EarthCube mission.</td>
</tr>
<tr>
<td>EarthCube Scientists</td>
<td>Use EarthCube resources and services to conduct scientific research.</td>
</tr>
<tr>
<td></td>
<td>Publish scientific results &amp; curate data as needed.</td>
</tr>
<tr>
<td>EarthCube Developers</td>
<td>Develop technologies and services that can be integrated into EarthCube.</td>
</tr>
<tr>
<td>EarthCube Architects</td>
<td>Establish EarthCube requirements, framework and operational concept.</td>
</tr>
<tr>
<td></td>
<td>Develop information model (vocabulary, ontology).</td>
</tr>
<tr>
<td></td>
<td>Establish standards guidelines.</td>
</tr>
<tr>
<td></td>
<td>Ensure interoperability between EarthCube Building Blocks.</td>
</tr>
<tr>
<td>External Data Users</td>
<td>Use EarthCube resources and services for research, education, and decision-making.</td>
</tr>
<tr>
<td>Curator</td>
<td>Ensure data is properly captured in EarthCube compliant data repositories.</td>
</tr>
<tr>
<td>Data Owner</td>
<td>Responsible for producing the data. Concerned about its distribution and use.</td>
</tr>
<tr>
<td>External Data Facility</td>
<td>Responsible for archiving data at other agencies (NASA, NOAA, USGS, etc); interoperability with the EarthCube Cyberinfrastructure.</td>
</tr>
<tr>
<td>EarthCube Governance Committees</td>
<td>Responsible for generating and monitoring the governance for the system including data curation, access, use case priority, interoperability standards, etc.</td>
</tr>
<tr>
<td>EarthCube Office Staff</td>
<td>Responsible for maintaining the community involvement within EarthCube and communicating changes and how to use the system.</td>
</tr>
</tbody>
</table>
Use Cases

- Big Science – Discovery, Comparison, Provenance, Model & visualization
- Collaborative Science
- Dark Data Contribution
- Tools Contribution
- Data Documentation
- Models Sharing
- High Performance Computing and Storage Resources
- Real Time Data
- Physical Sample Curation
Drivers

- Transform and accelerate research and discovery by turning data into knowledge and enabling interdisciplinary data integration.

- Provide critically needed data, tools, and computational resources and frameworks for cross-domain scientific collaboration, analysis and with long-term geoscience software and data preservation, discovery and use.

- Provide a geoscience cyberinfrastructure and architecture that is scalable, extensible and sustainable.
### Frameworks

- **Zachman Framework** - For organizing stakeholder concerns and perspectives.


- **Reference Model for Open Distributed Processing (RM-ODP)** – For architectural patterns for distributed systems.

- **Open Group Architecture Framework (TOGAF)** – For managing the architecture.

- **Federal Enterprise Architecture Framework (FEAF)** – For classifying the architecture into architectural elements and viewpoints.


- **ISO/IEC 11179:3 Registry Metamodel and Basic Attributes specification** - Provides a schema for a metadata registry.
• Scalability
• Community Driven
• Open Science
• Interoperability
• Sustainability
• Distributed
• Data Model Driven
Data Provider
Data Provider

Satellite Instrument Data Systems

Airborne Data

Agency Earth Data Archives

Other Data Systems (In-Situ, University)

EarthCube CI

EarthCube Repository

EarthCube Discovery
Benchmark

- Earth System Grid Federation (ESGF)
- Early Detection Research Network (EDRN)
- NASA’s Earth Observing System Data and Information System (EOSDIS)
Architecture Elements
Data Lifecycle

- **Original generation of data (from sensors, investigators, etc)**
- **Prepare data for use and submission into EarthCube**
- **Maximize information throughput against available bandwidth**
- **Supports the capture and validation of data into EarthCube**
- **Provides overall data management services for the data in EarthCube**
- **Enables discovery, access and distribution of the data**
- **Enables the analysis of massive, distributed heterogeneous data**
- **Provides a platform for integrating analytics with rendering and understanding the data**

Diagram:

1. Data Generation
2. Data Curation and Preparation
3. Data Transport
4. Data Ingest
5. Data Management
6. Discovery, Access & Distribution
7. Data Analytics
8. Visualization
Information Model Context

Information Model

Common (DC, ISO 19115, etc)
- Investigation
- Instrument
- Software
- Image
- Spectrum
- Time Series
- Data Set
- Granule
- Model Results

Atmosphere
- Oceans
- Hydrology

Integrated Information Model Knowledge Base

Used to Create
Validates

Schema

Product

Information Object (example)

Describes

Extract / Translate

HTML, DocBook
Config Files
RDF/SKOS
JSON
RDF/OWL
UML/XMI

Documentation

Software
Example Instantiation
Thank You

Questions?
EarthCube Conceptual Architecture Discussion

The controversial bits...
THIS IS A DISCUSSION.

Please Talk.
Stakeholders

- Do we have the right stakeholders?
- Do they overlap at all? Too much?
- Are they useful to provide use cases and personas that help drive the system?
- Are we missing key stakeholders?
## Stakeholders

<table>
<thead>
<tr>
<th>Stakeholder</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSF Program Managers</td>
<td>EarthCube Scientists</td>
</tr>
<tr>
<td>EarthCube Developers</td>
<td>EarthCube Architects</td>
</tr>
<tr>
<td>External Data Users</td>
<td>Curators</td>
</tr>
<tr>
<td>Data Owner</td>
<td>External Data Facility</td>
</tr>
<tr>
<td>EarthCube Governance Committees</td>
<td>EarthCube Office Staff</td>
</tr>
</tbody>
</table>

1/7/16
<table>
<thead>
<tr>
<th>Federation</th>
<th>Sustainability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standards</td>
<td>(Data) Model-Driven</td>
</tr>
<tr>
<td>Extensibility</td>
<td>Scalability</td>
</tr>
<tr>
<td>Provenance</td>
<td>Security</td>
</tr>
</tbody>
</table>
Standards...

- We do not advocate a particular standard...
- Our Conceptual Architecture emphasizes fully defined and self-contained data rather than prescribing standard(s).
- EarthCube’s heterogenous data, applications, and systems appear to justify possible increase in complexity.
- Common models and representations *should* be used.
EarthCube Software Lifecycle Processes

Technology Planning

Software Development

Release
Research Software Lifecycle Processes

1/7/16

Technology Planning → Software Development

Software Versioning → Software Search and Distribution

Algorithm Search and Distribution
Software Lifecycle Processes

- We place an emphasis on software versioning, discovery, etc. for Research Software. Should we treat “EarthCube proper” processes the same way?

- What about discovery and distribution?
Metrics

- Use Examples:
  - Product Searches
  - Products Downloaded
  - Services Accessed
  - Publications Cited

- Quality Examples:
  - Ingestion speed
  - Search Response Time
  - User “conversions”
Metrics & Conceptual Architectures

- Is this the right place to advise / mandate metrics? (e.g. we’re not doing this for standards)
- Should we be specific or just provide categories?
- Do we go so far as to "mandate" it for EarthCube components / building blocks / etc?
Places we haven’t expressed an opinion

- Cloud vs. on-premises hosting
- Data location (hosted vs. distributed)
- Compute location

Should we?
## Best Practices

<table>
<thead>
<tr>
<th>Common Software Stack</th>
<th>Common Data Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Interfaces</td>
<td>Service-Oriented Architecture</td>
</tr>
<tr>
<td>Decoupled Storage, Compute, and Data Management</td>
<td>Federated Search</td>
</tr>
<tr>
<td>Analytic Services</td>
<td>Visualization</td>
</tr>
</tbody>
</table>
Misc Questions

- How do we make this real?
- What’s the next thing you need to make EarthCube more valuable to you?
- How can the Conceptual Architecture effort help you get there?
Our Next Steps

1. Solicit Reviewers for Conceptual Architecture Document (NOW!)
2. Incorporate feedback and review comments
3. Write actionable recommendations and incorporate into final Conceptual Architecture
4. Prioritize and Deploy Key Architectural Components
We need reviewers!

Please contact Emily Law if you’re interested.

Thank you!