

## Site Visit Report

Project Title: **Collaborative Research: Integrated Arctic Data Management Services (IADS) to Support Arctic Research**

Award ID: **1016048**

Principal Investigator: **Serreze, Mark C.**

### Findings:

The National Science Foundation (NSF) convened a site visit panel consisting of the signees of this report to review the project, which started in July 2011, and examine its future plans. The project is aimed at designing and developing Integrated Arctic Data Management Services (IADS) to address the data management needs of the Arctic research community funded by NSF. The PI and his collaborators proposed to identify and catalog all data produced by NSF Arctic investigators and provide the assistance and tools needed to meet the requirements for publishing the data and associated metadata. They also proposed that the data would be well documented to guarantee preservation and to promote interdisciplinary reuse. The stated objectives of the project, quoted below from the proposal, are as follows.

- *Identify and catalog all data produced by OPP/ARC investigators.*
- *Assist Principal Investigators (PIs) and provide tools for data providers to meet NSF requirements to publish their data along with necessary metadata and documentation.*
- *Ensure that data are well documented for adequate preservation and broad, interdisciplinary reuse.*
- *Ensure that data are properly archived in accordance with international standards.*
- *Enable broad discovery of NSF and other Arctic data through diverse international and interdisciplinary portals, including the new IADS portal.*
- *Provide user-driven data integration services and products to enable the diverse data integration and reuse necessary to meet SEARCH objectives.*
- *Provide support services and help-desk functions for data providers and users.*

The panel unanimously agrees that a system meeting these stated objectives will be an invaluable contribution to earth system science. We wholeheartedly endorse the spirit and intent of this project, with further remarks herein intended to address what we perceive as *process challenges* based on the site review that took place June 4-5, 2013. The panel was provided with several documents, including the original proposal, all available annual reports and a brief summary of the previous site visit report. The PI,

10 June 2013

Co-PIs and staff made a series of presentations to the panel to provide an overview of ACADIS, the project's highlights, and demonstrations.

In the afternoon of the first day of the review and after most of the presentations on the agenda were completed, the site visit panelists met and prepared a set of questions for the management team to answer either in writing or in a presentation the following morning. The major concerns and recommendations were handed to the PI in writing, and the panel proceeded to discuss these comments with the PI, Dr. Mark Serreze, and the co-PI, Jim Moore. The PI requested that the document, along with any remaining questions, be sent via email to one of the ACADIS staff members. When the email was sent, the second page of questions from one of the panel members was inadvertently left out. The missing page was distributed the next morning after its absence was noted, and the panel requested that the ACADIS team provide its written response the next day (6/6/2013). This is included in the report as Appendix 1.

The panel was given several briefings, including demos on the second morning and a response to many of the panel's questions. These responses provided useful details that helped explain several aspects of the project. In particular, this showed that the ACADIS Gateway had solid and well-understood software architecture. Further, the different interfaces of Gateway and Arctic Data Explorer (ADE) were demonstrated. However, no overall system architecture flowing from requirements was presented, and the approach to the federation currently addressed by both Gateway and ADE was not clear to the panel. Useful usage statistics were presented, but this was not related to an evaluation of the "problem to be solved," so we did not understand what fraction of the data had been archived, what data should be archived in ACADIS, and how much is—and should be—archived in other repositories. This again relates to the federation issue. We note that some of these difficulties appear to stem from the loose coupling between the NSIDC and NCAR/UCAR work.

The main issue, which was first identified by the NSF panel that reviewed the original ACADIS proposal, is a lack of requirements and the translation of these requirements into a system-level design and project plan. The project has been pursued as a suite of largely independent and in some respects parallel activities, some of which appear to be motivated by needs related to other activities pursued by the contributing organizations. In other words, *missing* from the project is a comprehensive statement such as, "By such a date, ACADIS will provide the community the following set of capabilities: A, B, C in relation to existing, in process, and future data collections". Such a statement would in turn be supported by a coordinated system design and a corresponding system engineering process. The site visit panel is of the opinion that this deficiency is due largely to the distributed management of the project. It is vital that the management team understand this deficiency and work to review the requirements and clarify their project plans and project management. In addition, the panel recommends that an

experienced person be assigned (hired if necessary) to coordinate project activities and manage day-to-day operations, as was originally proposed.

We provide additional detail regarding these concerns in the following sections; as previously stated, Appendix 1 contains the project team's written responses to the set of questions posed by the review team. These answers are included for completeness.

## **Issues, Concerns and Recommendations:**

### **Project Management and System Engineering**

The panel identified several major concerns regarding the management of the development activities at the ACADIS project level. In summary, there is no integrated system architecture, no written requirements, and no configuration management at the project level. That is, there are three development activities (Gateway, ADE, Rosetta) that are run as independent projects; Gateway and ADE are largely redundant and none of the three are based on an overall requirements analysis.

The lack of project-level requirements means there is no accountability for compliance testing and design articulation to system engineering control. This is evidenced, for example, in the lack of a common metadata schema and controlled vocabulary, as well as in the existence of two separate software development activities that are not coordinated by a common set of requirements or, in fact, any clear documented requirements at all. This makes it difficult to impossible to define and control a design or design changes against a commonly understood set of requirements and compliance testing.

### **Software Engineering**

The software engineering approaches within Gateway and ADE, independently and in isolation, were reasonable and appropriate to current best practice. Both Gateway and ADE use the Agile/scrum development approach. Each appears to be individually effective. We heard in detail about the ACADIS Gateway architecture, which is built on modern Java enterprise subsystems; Rosetta and ADE also appear to use state-of-the-art components. However, we did not hear details about the evaluation process that led to the final decisions; for example, technologies like iRODS, NOSQL and semantic web ideas like SPARQL could be useful.

### **User Services**

The panel was encouraged by the effort put into user services. The flow of data coming into the ACADIS portal is moderately steady, although the numbers are small given the financial outlay. We encourage the project team to assign responsibilities amongst the ACADIS staff along logical lines both for accountability purposes and so that PIs know

where to go. When the committee asked questions about user services responsibility during the meeting, the response was that everyone had such responsibilities. We feel that this is much too flat to be effective.

Both the panel and the project would like the ingest of data to be as painless and quick as possible for a PI so that more data are preserved and the PIs return. The Data Submission tool developed under CADIS and enhanced under ACADIS is excellent, and the panel was encouraged to see increased use of the tool by the ARC community. While many investigators will prefer web submission and communication via email, we feel ACADIS should have a phone number for answering questions and making smooth initial contact. There are sufficient full time curators to allow the reasonable manning of this service.

We encourage a more aggressive approach to search out and find data providers. Once all of the active ARC projects that will generate data are identified, the staff assigned to user services should systematically contact the PIs, determine if they are producing data that should be archived, determine the correct archive and come up with a simple plan for preservation.

### **Rosetta**

Rosetta looks like a very powerful approach that has been needed for a long time and could go very far in helping to unify the earth science community around the NetCDF framework. Nonetheless, it is unclear why this is being developed under ACADIS. The presentations clearly stressed the fact that the investigator community did not want anything to do with NetCDF. While this may be collectively short-sighted, it is understandable that individual PIs would not want to have to learn to write NetCDF. A service like this would be a great contribution, but should perhaps be developed under base NetCDF development funding or by other funding. It could certainly contribute to the long-term goals of ACADIS, but in the face of the more fundamental deficiencies, it seems like an inappropriate use of resources for this project at this time. Furthermore, the spectrum of Arctic program-related data formats is much broader than the current Rosetta capabilities support (ascii and excel tables). As it will be difficult to accommodate the full spectrum of diversity, the usefulness of Rosetta for ACADIS needs may be limited.

### **Data Curation**

The success of ACADIS will ultimately depend on the extent to which the broad Arctic community trusts and feels “ownership” of the system. Developing a strong relationship with your community of data creators and consumers is essential to fostering this sense of ownership. The data curation services ACADIS is developing will play an important role in this. Creating and maintaining direct contacts with PIs to discuss what data they

have acquired, what data products will be generated and how they document their data to do their science will help ensure future contributions of high quality datasets. These direct contacts will also help ACADIS optimize the system for their data creators while also enhancing the visibility of ACADIS within the Arctic science community. The committee strongly encourages ACADIS to continue its existing efforts in this area.

The impression of the committee from the presentations is that the primary focus thus far has been in the development of Data Management Plan *assistance* which, while laudable, should evolve to focus more on DMP *execution* in funded projects. Other efforts have developed DMP tools that involve minimal staff time, and these may be useful models for ACADIS to consider for their future developments. Here again, an over-arching point is the question of cost-benefit resource management as it relates to system design and desired user-driven outcomes.

Experience from other projects indicates high return from time invested in contacting PIs of funded projects; identifying contacts early in a project can help ensure better data documentation during the data collection phase and more reliable data submission.

### **Data Preservation**

The core goal of ACADIS, to ensure the preservation of all data generated under the NSF-ARC program is ambitious but attainable with the current support provided to ACADIS and the earlier work done by the CADIS team. There are three broad classes of data: 1) data that are housed in other existing expert archives; 2) field data from real-time, data-logger type sensors that are not already archived in an existing facility; and 3) “special” data ranging from social science data; to analytic, derived or synthesis data products; to the high-sample rate devices and imagery that were identified in the presentations. The first type of data has not yet been accounted for in the ACADIS design. Data set records from existing archives will need to be harvested without requiring redundant registration from Arctic PIs.

The second and third types of data will be diverse and heterogeneous in format but may be small in size ( $<10^{12}$  bytes). Some data types and formats will likely be unique and may require special handling (for example social science data with privacy issues). However, there are existing disciplinary repositories appropriate for some ARC data, and **these existing resources should be leveraged wherever possible.**

Metadata standards are lacking at the data set level for all data types, and these disciplinary repositories are best suited to deal with the challenge of developing the appropriate metadata standards. While metadata for all ARC programs clearly must reside within ACADIS, the actual data files can reside in external disciplinary repositories as appropriate. ACADIS does not need and arguably should not be the archive for all ARC data.

The panel recommends that ACADIS develop a catalog service that links to external repositories. Rather than duplicate data archiving services provided with other systems, ACADIS should focus its archiving resources on those data types without other homes. User support services should include directing investigators to these other repositories as appropriate for archiving their data (including in the DMP).

### **Web Service**

The traditional method of conducting research based on a Consumer data transaction is as follows:

- *Go to the web site*
- *Browse around*
- *Find the dataset*
- *Download it*
- *Reformat it*
- *Load the reformatted data into a program*
- *Proceed with the analysis*

The problem with this approach can be illustrated in few words: Do this 17,000 times. The Web Service/API approach breaks down this “wire” in a different way:

- *Write a program*
- *Include specific access to source data through the API.*
- *Run the program to do the appropriate exchange 17,000 times.*

We recommend that Service API's be provided for all ACADIS resources.

### **Cultural Sensitivity of Archives**

When interacting with Arctic indigenous peoples, sensitivity to cultural issues is paramount. The ACADIS resources should therefore present culturally-sensitive interfaces. The project should explore the experiences of AIHEC (American Indian Higher Education Consortium) and NMAI (Smithsonian's National Museum of the American Indian) to further educate team members on this issue.

### **User Scenarios**

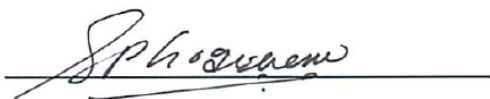
User scenarios will help immensely and across the board in near-term continuation of the project. In particular, beyond the obvious “hermetic” use cases (where the researcher needs nothing more than what is available from ACADIS), there should be a clear view of how NSIDC and ACADIS will be interoperable in a meaningful fashion with other data centers.

10 June 2013

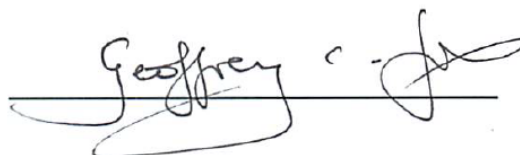
## ACADIS NSF Site Visit and Review – Panelists

Day 2 at NCAR/CISL Vislab, 8am-12:30pm

Prasad Gogineni (Chair)

Handwritten signature of Prasad Gogineni in cursive script, written above a horizontal line.

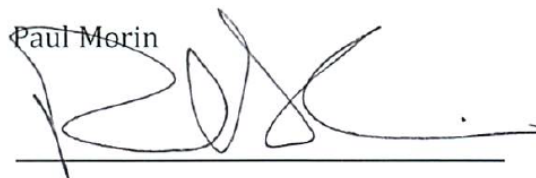
Geoffrey Fox

Handwritten signature of Geoffrey Fox in cursive script, written above a horizontal line.

Suzanne Carbotte

Handwritten signature of Suzanne Carbotte in cursive script, written above a horizontal line.

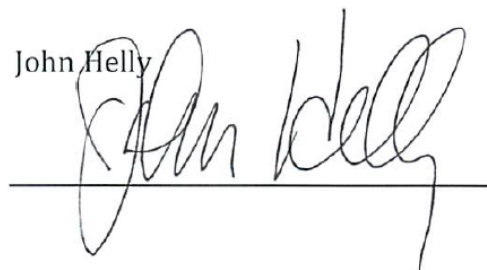
Paul Merin

Handwritten signature of Paul Merin in cursive script, written above a horizontal line.

Rob Fatland

Handwritten signature of Rob Fatland in cursive script, written above a horizontal line.

John Helly

Handwritten signature of John Helly in cursive script, written above a horizontal line.

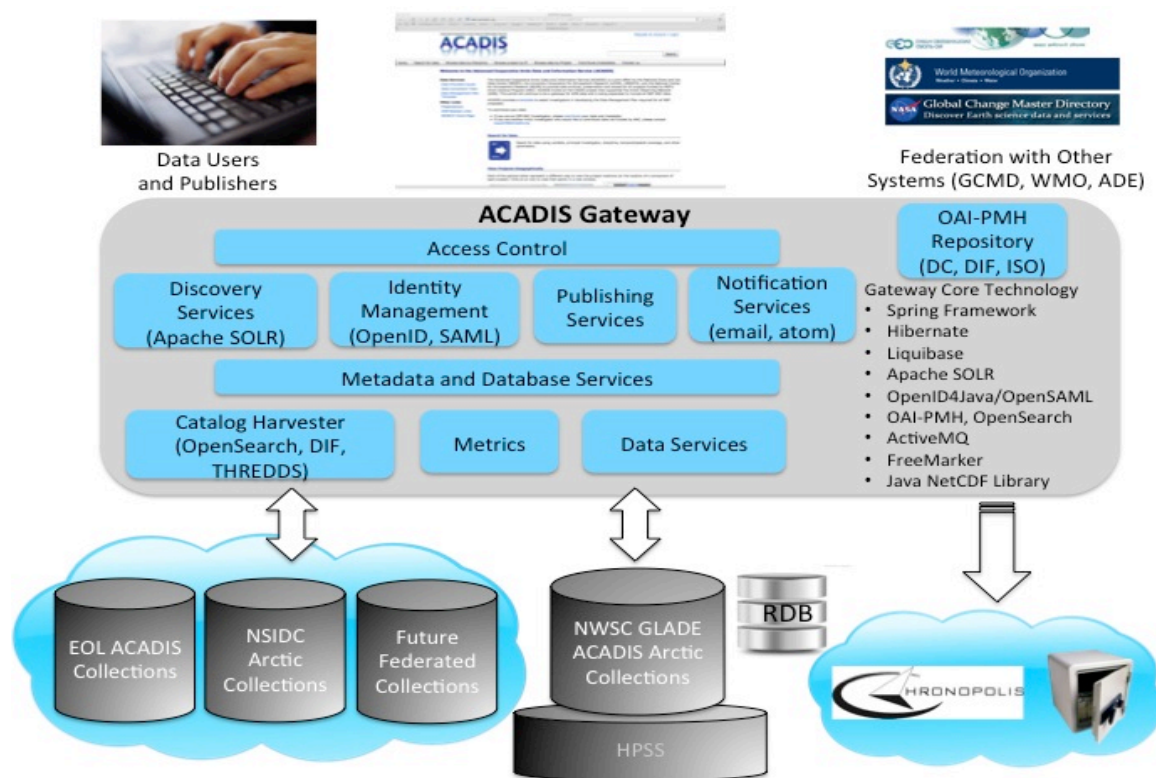
## Appendix 1

ACADIS team written responses to NSF review panel – 2013-06-06

1. We did not understand the data architecture. This could have components such as federation, repository system for new datasets, existing repositories and a design for metadata that spans new datasets and existing repositories. Is there a system architecture/diagram for the data architecture?

The ACADIS Gateway, (<http://aoncadis.org>), was envisioned as a shared, central clearing house for NSF Arctic and related data from other sources for NSF researchers. It has evolved substantially over the last several years in terms of holdings, capabilities, and its ease of use. At present, the ACADIS Gateway provides services for data-publishing research teams and individual PIs, data consumers, data curators and support personnel, and NSF or other staff interested in verifying publication of research results. In addition to the publishing capabilities, the Gateway provides browse and search capabilities, authentication and group authorization, federated identity, tools for curators and support staff, federation of catalogs to other systems, federated search locally of other systems, basic support for digital data citations (via DOI's), and digital preservation support via Chronopolis.

The following figure provides a high level view of the ACADIS Gateway's architecture.





10 June 2013

On the right-hand side of the figure we describe Gateway Core Technology and some of the components the system is built upon. It's a very modern and practical software stack relative to the Enterprise Java world, and allows us to take advantage of existing and emerging technology supported by global, community software engineering efforts. We integrate rather than invent whenever possible.

Can you clarify relation between the Arctic Data Explorer and the ACADIS gateway? They both seem to have federated search interface.

This was discussed at length in the morning session.

How many datasets are registered within ACADIS? – Can they provide a chart of growth of datasets over ACADIS period? Also over the CADIS period could provide useful perspective.

The following charts show self-publishing activity and cumulative data collection count since November of 2009. The charts below do not reflect special requirement datasets of which there were 84 in the first year and 62 in the second year.

What is the long term plan to rationalize various interfaces and modes of access for Arctic data with the larger holdings of NSIDC?

As discussed in the morning session, the Arctic Data Explorer is based on the same infrastructure as the NSIDC NASA DAAC search.

How could the cataloging and data ingest be accelerated?

Accelerating metadata cataloging and data ingest has both a technical and a social aspect. Enabling data providers to quickly and efficiently publish data products to the Gateway via the self-publishing interface will lower barriers and accelerate the process. Year 3 plans include dataset management tools for easier data re-organization, easy to understand service interfaces for automated data ingest, automated metadata extraction where possible and integration of tools such as Rosetta to streamline data publication and still retain rich metadata capture. Proactively engaging and drawing in eligible data providers toward ACADIS services will likely yield the strongest acceleration effect. To date we've relied on announcements in newsletters (ArticInfo/WTA), word-of-mouth, introductions from NSF PM, and flyers at conferences such as AGU. These approaches have not been effective at the level we desire. We gratefully acknowledge suggestions made during our panel review, including direct and proactive communication with each active NSF active award recipient(s) which we will pursue as a priority.

Can you outline how better contact with PIs can be made?

This will be accomplished with further development of the ACADIS Gateway plans and tasks along with visibility in the community as part of meetings, demos, and additional help from NSF providing guidance to the funded OPP PIs and collaboration with other Arctic researchers. We hope to engage early during the grant proposal process and target communication to PIs of funded awards with the help of our NSF PMs.

How is the metadata exposed without going through the CADIS interfaces? Web services?

Please refer to the ACADIS Gateway Architecture figure presented above. In the upper right corner of the figure the Panel will note a box labeled OAI-PMH, which expands to the Open Archive Initiative Protocol for Metadata Harvesting. We expose an OAI-PMH web service that can deliver Dublin Core (DC), NASA's Directory Interchange Format (DIF, used by GCMD), and ISO 19115/19139 WMO Profile V1.1. In the coming year we expect to also provide ESIP OpenSearch, and potentially others as requirements arise from community needs. As mentioned to the Panel during our in-person meeting, these are open feeds i.e. not protected by security or firewalls. We also note that the Panel recommended that we advertise these links so any group can consume them, and we agree that this is a really good idea and now have it in our Agile backlog for the Gateway. Metadata are available through the GCMD and the Arctic Data Explorer interface and ESIP OpenSearch output.

To communicate ACADIS motivating concepts and scope we would suggest providing some exemplary User Stories. (At least three) such stories would establish by example physical and temporal context, use cases that cover various data types, standard research activities, incentives for researchers, scalability to the community, and other central data system features.

The ACADIS precursor, CADIS, took this approach early on to better frame discussions and development. We plan to revisit this earlier work as well as further leverage our Advisory committee for user stories and direction. Questions remain as to non-'typical' datasets and handling of datasets with special requirements.

## **1) Project Management**

1. Schedule: what is the schedule for development and operations?

Given that all partners have other projects, schedules for development and operations for ACADIS are coordinated within each organization to meet all deadlines and deliverables. Rolling 3 month priorities are discussed monthly among the Sr. Management team. Where dependencies between efforts exist, schedules are coordinated in the monthly Sr. Manager meetings, in the subgroups, in the weekly curator/CISL development meetings, or between involved personnel as needed.

ACADIS is operational and continues to develop; further developments are planned based on user input and ADAC guidance.

The gateway has been in production operation throughout the full span of the ACADIS project, having become operational during the first CADIS project. The CISL development team uses an Agile Scrum software engineering methodology. We utilize an ordered backlog with the goal of delivering the highest value features first. We regularly revisit the backlog to ensure the backlog contains the most relevant items in relation to the current realities of the project. Our team works on items from the backlog in 2 week sprints. Each sprint concludes with a sprint review which is intended to result in a potentially releasable software increment. Many sprints conclude with direct production deployments, and we are currently averaging a little better than one production deployment per month (this is a vast improvement over where we were two years ago). As the gateway is an operational system we are very focused on evolutionary development approaches to ensure we can always deliver features to production. Our approach is to strive for delivering a steady value stream over the life of the project.

2. What is the project-level architecture and how are functions and interfaces identified and documented?

Please refer to the Gateway Architecture figure above. This shows some, but not all, of the interrelationships across ACADIS technology components. In the upper right corner, the harvesting of ACADIS Gateway metadata catalogs by the Arctic Data Explorer is accomplished via OAI-PMH as a conduit for DIF records. EOL Special Requirements Data Collections (SRDC's) are harvested by the Gateway via THREDDS catalogs. We have not yet designed the Rosetta integration so it's not indicated in the figure; there are several possibilities and/or phases and defining this is a planned Year 3 activity.

Regarding the Panel's good question about how functions and interfaces are identified and documented, we strive to leverage community and international standards whenever possible. The architecture figure calls out several of these e.g. OpenID, the Security Authorization Markup Language (SAML), OpenSearch, OAI-PMH, ISO metadata, DIF metadata, Dublin Core metadata, THREDDS, Apache SOLR for search/discovery, and so forth. All of these have extensive documentation on the web, and sometimes in books. The Gateway currently has no non-standard interfaces or protocols, and we intend to keep it that way as much as possible.

3. What other data center approaches have been evaluated to arrive at this design?

Neither CADIS nor ACADIS initiated a fresh data gateway design/build effort. Instead we first leveraged NCAR's Community Data Portal (CDP) for CADIS, and then migrated

to the Science Gateway Framework (SGF) upon which the earthsystemgrid.org (ESG) site was also built. Requirements and specifications for the SGF/ESG grew out of the international climate model data community, and reflect input from multiple agencies. The SGF was extended and continues to be enhanced in a number of ways to satisfy the needs of the Arctic science community. We basically have heavily leveraged large investments in cyberinfrastructure that had been proven in operation over a period of years.

The ACADIS Team is constantly interfacing and talking with other Data Centers as part of ACADIS as well as other projects with distributed archives. ADE is also investigating metadata interoperability as part of its design and development.

#### 4. What is this design?

The Gateway Architecture diagram along with several of the paragraphs above should provide a good indication of the overall functionality, the architecture and major functional components, interfaces to external systems, general design philosophy (e.g. community standards, open source software), and software engineering methodology.

## 2) Software Development

#### 1. What is the approach to software engineering at the project level?

Within the ACADIS Scrum approach we regularly review the backlog items to ensure it is relevant to the current project realities. Coupled with 2 week sprints we have many opportunities to coordinate development activities with the other groups in the collaboration. Through year 2 there have been a few integration points, most notable exposing the OAI feed for harvesting into the ADE, GCMD, WMO-WIS.

Given the low volume of integration points the architecture group has met a few times to review the OAI integration issues through year 2. This is expected to change significantly for the current year 3 work plan. We are expecting to work on federated authentication and authorization in addition to Gateway-Rosetta integration. This will likely require federated authentication, authorization, and API development and coordination. As the Architecture SubGroup is slated to lead this effort we anticipate convening the relevant engineers much more formally to address the architectural and engineering needs and efforts.

The approach thus far has been to drive Rosetta development based on use cases, requests, and usability feedback from the community. Unidata uses Jira for bugtracking and feature requests. Bug reports and feature requests are grouped into release targets, which drive our development roadmap. Code is tracked and shared using the distributed version control system Git, paired with the online git hosting service Github.

10 June 2013

Maven is used to handle build and dependency management for Rosetta, and the resulting Rosetta artifacts will be available through the Unidata maven repository (<https://artifacts.unidata.ucar.edu/content/repositories/unidata-releases/>).

EOL isn't actively engaged in the ACADIS Gateway development. That said, the EOL software engineers designed and integrated the original metadata input form and database for CADIS. EOL maintained the metadata database until CADIS became part of the CISL Gateway architecture, about halfway through the 3 year CADIS period of performance.

EOL collects metadata and maintains a database, and has developed a metadata input form that is based on the ACADIS profile with added extensions. The EOL arctic metadata profile is a superset of the ACADIS one, with additional fields to handle metadata for biological and special requirements datasets. EOL has adopted agile programming methods, e.g. rapid turnaround and code sprints. We forego formal scrums, generally, because we are a couple developers who keep in close touch with each other. We code in Grails and Groovy to Java byte code for the JVM.

2. How are the code-bases for the gateway and the Arctic data explorer maintained?

ADE is a JavaScript single page app whose code is maintained in a Git repository at NSIDC. Code changes are 100% test driven. The GI-Cat metadata broker app is a Java app hosted in a set of official SVN repositories maintained by CNR. NSIDC maintains a local clone (using Git) of these repositories, develops, integrates and tests changes on this local copy, and contributes completed increments of functionality back to the CNR repos. ADE/GI-Cat system changes are covered with automated acceptance tests, and subject to automated test and deployment within a continuous delivery pipeline.

CISL maintains the gateway and related components in an institutional Subversion repository. This is a centrally managed service used by many groups within NCAR. Support groups within CISL manage and maintain the service with an excellent track record for service and reliability. Upon request we can add external users to the repository with very little overhead. As a managed service this doesn't require any resources on our part to maintain.

There are several efforts underway to evaluate what the next generation SCM system should be for the organization. We are looking at both GitHub and Stash, an Atlassian product comparable to GitHub.

Subversion is part of the suite of application life cycle tools in use by CISL

- Subversion – Software configuration management

10 June 2013

- JIRA – Issue tracking, Agile planning, and user support tracking
  - Bamboo – Continuous integration build server
  - Crucible/Fisheye – Online code browsing and peer code review
  - Nexus – Software repository
  - Sonar - Code quality management
3. how are bugs tracked and sourced from user-support?

Incoming bugs, identified through internal testing or users, are submitted through the support@aoncadis.org email address and logged in a central ACADIS Zendesk instance. Future plans include implementing the Zendesk web forms for additional communication pathways. The ACADIS Zendesk instance is connected with the CISL JIRA system to automatically represent bugs in the Gateway development backlog. CISL uses JIRA to track bugs. Bugs are tracked in the product backlog along with user stories. CISL also uses JIRA to track user-support requests using a Kanban board to visualize the current state of all support requests. We have integrated the ACADIS Zendesk help system to our JIRA instance so gateway specific tickets can get automatically routed into JIRA. We have also reviewed support requests to look for patterns of user difficulty to help identify areas where we can improve the user experience.

The ACADIS Zendesk system is also connected with the separate NSIDC instance of Zendesk. Arctic Data Explorer code bugs are routed to the development team through the NSIDC Zendesk system, where the Agile team Product Owner prioritizes the bugs against the existing backlog of work. Pivotal Tracker is used as a backlog management tool, including tracking and prioritizing bugs. If the bug is determined to be in the GI-Cat system code, a decision is made between the NSIDC and ESSI-lab team as to a best fix and which team should do the work.

The ACADIS Zendesk instance is additionally connected to Unidata's JIRA implementation, where bugs and feature requests for Rosetta are tracked. Bugs identified by non-ACADIS users can be routed through Unidata support (support-rosetta@unidata.ucar.edu) and items are entered into Jira as needed. Support requests submitted to Unidata are archived and made available to search engine crawlers for public availability.

4. how are interfaces defined and documented?

ACADIS Inter-system interfaces are standards-based web service interfaces, e.g. OAI-PMH/DIF, ESIP OpenSearch. These system interfaces have been developed by external organizations and leveraged for our purposes.

At this point in the project we have utilized existing community defined interfaces and protocols such as OAI-PMH, OpenSearch, DIF, and THREDDS. The result is that we can rely on existing documentation for the protocols and interfaces developed by others.

5. how are development releases tested and released to operations?

We perform testing and QA processes during each sprint cycle. This process includes automated unit tests as well as manual user testing. At the end of each sprint the code is viable to put into production. A benefit of short 2 week cycles is the amount of possible change is relatively limited compared to a long release cycle. As such we have been able to remove the large, and expensive, full regression test cycles while significantly reducing our overall defect rates.

Our full release cycle takes advantage of three distinct deployment environments. During the sprint we have a QA environment for testing the current state of Gateway. This is updated often, sometimes multiple times per day, and is available for demonstration, review and testing.

Once a release has been cut it gets deployed to our staging site. Final deployment verifications can be made here and this also serves as a production mirror for any testing and validation.

Finally, we have a production environment where the gateway releases are deployed. The engineers that developed the code are also responsible for operating and maintaining all three environments. This gives the developers a very direct feedback loop into their work.

Throughout the process we have a continuous integration build server (Bamboo) running. We do automated nightly builds and the official release builds through this tool to ensure consistent and repeatable builds. All of the builds deploy artifacts to a publicly available software repository where they can be download by the community if desired.

The Arctic Data Explorer is using a continuous delivery system for the ADE/GI-Cat codebases. Upon commit to version control, unit tests and other code quality tools are run for all affected projects. Successful completion triggers deployment cycles for a production-like integration environment, upon completion of which automated acceptance tests are run. Once a feature is completed a push-button step is used to deploy (and test) to a QA environment for evaluation by the Product Owner. The PO approves another push-button deployment of completed features to a staging

environment, from where NSIDC's operations team evaluates the system and recent changes, performs a push-button deployment into production, and ensures continuity of service.

6. how are project resources backed up?

Project resources including the ACADIS Gateway metadata database and all dataset files are backed up daily to the NCAR HPSS tape storage system. These backups are retained for a 6 month period and then aged off the HPSS as needed. CISL engineering software life cycle tools, such as subversion source code repository, JIRA issue tracking system are backed up daily to the HPSS in a similar process. [

We have recently archived a snapshot of all ACADIS data collections and the entire metadata repository into the Chronopolis digital preservation system. Chronopolis is a carefully managed dark archive with resources at SDSC, Univ. of Maryland (UMIACS), and NCAR. The end result is a copy of all ACADIS resources held securely in three geographically distant Chronopolis nodes. In the future we intend to automate the process of archiving to Chronopolis and to pursue incremental additions rather than full snapshots. The overall idea is to reduce the probability of losing any ACADIS resources to near zero, and this only in the event of a national-scale catastrophe.

The ADE and associated GI-Cat source code and configuration management repositories, as well as user-facing documents (Data Management Plan template, etc) are backed up with nightly incremental backups and quarterly full backups, both to LTO4 tape, with a retention of 6 months (i.e. minimum of two full backups). NSIDC does not store other resources for these projects.

7. how are user registration and authentication being done?

User registration and authentication are handled by Identity Management (IDM) components within the gateway. At this point in time authentication is username and password, however the Gateway IDM components support both OpenID and SAML protocols.

Broader federation needs in year three are expected to drive the need to utilize the OpenID and SAML capabilities in the Gateway. Additionally we are investigating the use of OAuth and XACML in enhancing the federation capabilities of the ACADIS system.

8. How are users apprised of updates, anomalies with their data, and tracking of downloads by data author being handled?

The ACADIS Gateway tracks individual file downloads in the relational database. The related dataset (and containing project) can be identified and therefore downloads by dataset (with related author) and project (PI, Co-I, etc.) can be aggregated and



reported. At this time, ACADIS Gateway based dataset files are freely available and authentication is not required for access. As a result, the identity of the downloader is not known (beyond IP location) and notification of dataset changes or other issues cannot be readily accomplished. Year 3 plans include providing feeds (such as Atom/RSS) for individual datasets which would allow users to opt into notifications of metadata and file changes. We're considering requiring authentication for download (or some other mechanism) to capture a user's email address for proactive notifications.

For the Special Requirements datasets, we send e-mails to those who have ordered data when datasets have been revised. When a dataset has been added to the archive, we send e-mails to the PIs and authors of the dataset and cc: the Data Manager.

### **3) Data and Metadata**

#### **1. What is the approach to metadata development?**

ACADIS established the Metadata SubGroup, evaluates the needs of multi-disciplinary datasets, and receives guidance from ADAC.

#### **2. Data and metadata workflow (e.g., ingest, review, metadata production, review, publication): what is it?**

There are two fundamental workflows for authoring dataset metadata and registering files: the self-publishing workflow supported by the ACADIS Gateway and the special requirements dataset workflow.

#### **3. Quality Control: how is it being done?**

Metadata quality control is enforced primarily via the ACADIS Gateway self-publishing user interface. Input is constrained via controlled vocabularies, required fields are enforced and rudimentary format checking is applied to metadata element types such as dates and geographic extents.

Quality Control of the data is normally the responsibility of PI or the data provider since they are in the best position to do so. But ACADIS Data Curators do check for metadata completeness and correctness. If problems are discovered, the PI or data provider is contacted.

#### **4. Anomaly Detection and Errata Reporting: how are anomalies and errata reporting being handled?**

This is normally done as part of the "readme" documentation and feedback from the community to the PI or Data Provider.

10 June 2013

5. Data Publication: what is the metadata schema to the file level?

Rosetta uses the netCDF Climate and Forecast (CF) Metadata Conventions (version 1.6) when collecting file level metadata via the web-based interface.