A CUAHSI DataCenter for Hydroinformatics: Draft Specifications

Alva L. Couch
Tufts University
February 21, 2012

TABLE OF CONTENTS

INTRODUCTION

The purpose of the proposed project is to create what we will call a CUAHSI "datacenter for hydro-informatics." The motivation for creating this datacenter is to support access to and use of water data sources by academic researchers. Activities of the datacenter will include curation of water data and catalogs, as well as curation of data publication and access standards, and embodiments of these standards in data access and analysis software, including the existing CUAHSI Hydrologic Information System (HIS) open-source suite of software tools and services. Envisioned future activities of the datacenter include creation, distribution, and curation of high-quality data products targeted at hydrology and derived from data products in other domain sciences.

Scope of this Document

This document is not a complete proposal, nor does it attempt to scientifically justify the above missions of the datacenter. Instead, this document focuses upon the feasibility of accomplishing the above missions, the technical details of doing so, the design of the overall personnel and computing infrastructure necessary to do so, and the cost of that infrastructure.

This document is separated into three logically distinct parts: requirements, design, and implementation.

- The **requirements** section lists the overall requirements for the datacenter, including mission, relationship to prior work, and perceived high-level challenges of meeting the mission.
- The **design** section specifies how the datacenter will meet requirements via high-level capabilities, processes, and personnel roles.
- The **implementation** section describes policy decisions to be made in implementing the design, including a plan for overall sustainability of the datacenter.

REQUIREMENTS AND MISSION

The requirements of the datacenter include its mission, context, and a discussion of the technical and governance challenges of setting up such a datacenter.

Mission

The short description of the datacenter's mission is to advance the state of the art in academic research in hydrology and related disciplines by empowering academic use of time-series information sources. This mission will be accomplished by promoting sharing of time-series data, by developing data sharing standards, and by developing software based on these standards for data access and analysis. The DataCenter will focus on sensor data collected *in situ* at both fixed points and from moving platforms, although some laboratory data, such as aquatic chemistry data, can be accommodated. Coordination between the proposed center and existing Data Centers for gridded atmospheric and climate data (Unidata and NCAR), seismic data (IRIS DMC), and geochemical data (IEDA) is a recognized need and has been explored during the pilot project. The proposed center fills the niche for physical, chemical, and biological data (simply called "water data" throughout the rest of this report) that is currently unmet.

This center will serve the diverse academic research communities that utilize water data. The datacenter will transform existing research prototypes and standards for water data access and publishing into production-level services based upon these prototypes and standards.

The mission of the proposed datacenter's primary mission *to empower scientists to discover, use, and share water data* has five main components:

1. **Community governance** of datacenter activities.
2. **Standards** that foster information reusability and interchange.
3. **Curation** of water data and catalogs that conform to and realize those standards.
4. **Software** that embodies these standards and empowers research.
5. **Support** that empowers researchers to utilize data and software for scientific inquiry.

Thus the proposed datacenter is not just a data repository, nor is it just a software development effort: it is a coupled empowerment-centered effort that crosses the artificial boundaries between standards, data, and software to serve the specific needs of Earth and environmental scientists.

In detail, the datacenter will:

1. Develop, curate, document, and support community-endorsed standards and best practices for water data use, publication, sharing, and discovery (including WaterML and WaterOneFlow services).
2. Develop, curate, maintain, and adapt a community-governed set of water data access tools to the changing needs of academic researchers in hydrology and related disciplines.
3. Foster collaboration between hydrologists and software developers in creating, curating, and sharing new approaches to water data recording, publishing, discovery, and analysis.
4. Curate and maintain the CUAHSI catalog of water data sources, and ensure interoperability between CUAHSI catalogs and future federated data catalogs.
5. Develop, curate, maintain, and adapt methods for transforming water data from other data sources into formats suitable for cataloguing (e.g., by creating wrappers that provide WaterOneFlow services based upon data in other formats, or by creating, distributing, and curating derived data products of use in water research).
6. Partner with major environmental observatories in assuring that their data is accessible and useful to the academic research community.
7. Partner with vendors of sensor equipment to assure that sensor data is made available in a form usable by academic researchers.
8. Support academic researchers in publishing, discovering, and analyzing water data.
9. Develop services that curate, maintain, and publish research data for independent academic projects.
10. Assist larger academic projects with publishing and curating their own water data, in ways that maximize data sharing and usability by other academic researchers.
11. Interact with government agencies providing water data, to promote suitability and usability of data in academic research.
12. Foster community information exchange of data, software, and standards within the diverse research communities that utilize water data sources.
13. Identify challenges and obstacles in applying water data sources to academic research problems, coordinate community efforts in meeting the challenges, and represent the research community in communicating challenges and priorities to funding agencies.
14. Represent academic research interests to decision makers for government data sources (e.g., USGS, EPA) and data standards organizations (e.g., OGC).

One feature of the datacenter is a unique marriage between science and software engineering. The datacenter is governed by a community of scientists who determine needs, in partnership with software engineers and Information Technology (IT) engineers who provide for those needs. Standards, curation techniques, software products, and user support are all products of an ongoing dialog between science and engineering. For example:

- Scientists determine the need for a standard; software engineers analyze its impact and implementation issues.
- Scientists determine curation workflows; IT engineers implement the workflows.
- Scientists define new software features as designs or prototypes; software engineers incorporate these features into products.
- Scientists define needs for support; IT engineers provide that support.

A key attribute of the datacenter is a conscious interplay between cost and value. Scientists determine the potential for value; engineers determine the costs and contingencies. This balance between cost and value is necessary to developing supportable standards, catalogs, and software. A key goal of the datacenter is to obtain maximum scientific value for the cost by exploiting economy of scale.

Constituencies

The constituencies to be served by the datacenter include:

- **End-users** engaged in academic water research or related disciplines.
- **Data providers** who supply water or geoscience data for general use.
- **Software developers** who contribute software for water data discovery, mapping, modeling, and other academic research uses.

The datacenter provides the following value propositions for each constituency:

- **End-users** gain a single point of contact for expertise and mentoring in water data discovery and modeling, including instruction on standards for data publication, and protocols for data discovery and analysis.
- **Data providers** gain a single point of contact for expertise and best practices for publishing water data, including instruction and support for applying data publication and protocol standards to maximize the usability of data.
- **Software developers** benefit from the datacenter's unique combination of hydrology domain knowledge and software engineering capability. This unique combination assures that software contributions are incorporated into a high-quality software product that is more than the sum of its parts.

This last point deserves a detailed explanation. CUAHSI HIS is an open-source project that is – at this point in time – designed for and by hydrologists. CUAHSI HIS is important not just as a software resource, but also, as an embodiment of data sharing standards and best practices. The datacenter will preserve that unique quality of the project by adding engineering components to the existing prototype, including:

- Strategic re-engineering of parts of the existing CUAHSI HIS software suite for increased reliability and robustness in production.

- Designing and implementing comprehensive test and release management plans to increase the reliability and robustness of new releases of CUAHSI HIS.
- Providing development support to developer-scientists who contribute extensions to CUAHSI HIS, including email and phone technical support; integration testing and release management; and ongoing software maintenance services for the life of the extension.
- Providing engineering support to transform contributed feature prototypes into production features.
- Establishing and implementing curation workflows that ensure that datacenter data remains accurate, relevant, and useful to water researchers.

Thus the datacenter serves as a unique and innovative point of synergy between energetic scientist-developers and expert software engineers, empowers the hydrologists to make a difference in the overall community, and offloads much of the work of putting their ideas not just into practice, but also into production use.

Metrics of Success

The above plan is based upon serving the needs of the hydrology research community. Thus the metrics of success concern how well that community is served, including:

1. **Usage:** how many researchers actively use data from the datacenter?
2. **Citations:** how many research projects cite the resources of the datacenter in their publications?
3. **Subscriptions:** how many researchers are willing to pay for subscription services from the datacenter?
4. **Performance:** how reliable is the service, and how quickly can researchers find data of interest?

A good score on the fourth metric is crucial to good scores for the other three. In e-commerce, there is a "5-minute rule" that if a customer does not find what is desired in 5 minutes, the sale is lost. Here there is an analogous situation: the researcher or student should obtain some kind of "gratification" – usually finding data of interest – in the same five minutes. This strongly suggests the use of a powerful web-based client to enable searches, but also requires that the data sources be carefully curated. A search service that reports data services that are down or otherwise unavailable violates the 5-minute rule.

These metrics in some sense determine the priorities for implementation of the datacenter:

1. **Reliability** of services.
2. **Curation** of catalog and sources.
3. **Ease** of access.

Reliability comes first, followed by rigorous curation processes that avoid null search results and wasted time, upon which a streamlined web-based user experience can be based. To some extent, these three priorities can be pursued in parallel, though the success of efforts later in the list depend upon success of earlier ones.

CHALLENGES

Accomplishing the above missions requires rising to and addressing both technical and governance challenges. Technical challenges are those that have some technical solution, in the sense that implementing extra software or processes can address the challenges. Governance challenges are those that arise from diversity of needs within the user community.

Technical Challenges

There are several challenges in curating, cataloging, and searching water data that the datacenter must address to accomplish the above missions:

1. Water data are collected at many time scales, for many different reasons, and with varying regulation and quality control standards.
2. Published data change in quality over time, due to quality control processes that occur after data collection. Data are often published before quality control processes are applied.
3. Data publication sites differ substantively in mission, methods of data curation and publication, kinds and accuracy of recorded metadata, and data semantics.
4. Data and metadata at external sites can change without notice, thus invalidating catalog entries that describe them and/or services that depend upon them.
5. Data publication does not integrate well with other activities in pursuing water science.

CUAHSI has already played a part in addressing some of these challenges, including:

1. Community-governed standards for metadata (WaterML).
2. Community-governed standards for data delivery (WaterOneFlow services).
3. A reference implementation of standards for data delivery (HydroServer).
4. Technical support for researchers who wish to implement their own data publication services.
5. Wrappers that translate from existing non-standard to standardized data publication services.
6. A reference implementation of standards for data search and discovery services (HydroCatalog).
7. A GIS-enabled data discovery client (HydroDesktop).
8. Technical support for researchers who wish to implement their own data discovery clients.

The proposed datacenter will curate and maintain these existing mechanisms, but must also solve new and emerging problems by updating standards and developing new ones where necessary. For example,

- There is no standard for expressing data quality. The ODM standard expresses quality factors as user-defined annotations, while some services (e.g., USGS) provide data with quality tags with rigorously defined semantics whose values change over time as quality control procedures validate prior measurements.
- There is no standard for implementing data discovery services, though HydroCatalog provides an initial prototype of a possible standard.
- Wrapping and repackaging new data services as standards-compliant WaterOneFlow services requires semantic mapping from external to internal, standardized semantics for metadata. Previously wrapped and mapped services often change to provide new or different metadata as the services are maintained or updated over time, so that remapping is necessary. There is an ongoing need to check and re-validate these mappings over time.
- Termination of funding commonly results in the disappearance of valuable online data sources. There is a need to capture the contents of these data sources for archival purposes and to continue to make the data available via other means.
- There is a need to update hydrologic standards, reference implementations, and curated data to interoperate with proposed federated data catalogs.

Governance Challenges

As well as the above technical challenges, the datacenter will also face some challenges of community governance.

First, the governing body of the datacenter must choose an appropriate balance between:

- "Lowest common denominator" services that serve the largest communities of researchers.
- "Highest common denominator" services that support smaller communities in redefining the state of the art in research.

The tension between these two activities is already evident in current and past CUAHSI HIS activities, for example:

- The "HydroSeek" and "HydroExcel" search clients that aim to serve the search needs of the largest possible population of users with a simple interface.
- The "HydroDesktop" search client that serves advanced needs of researchers in correlating water data with maps and other geospatial data.

Experience has been that these two goals are independent of one another; each sub-community needs its own kinds of tools. One of the key challenges of community governance will be to achieve a balance between advancing the state of the art for the greater population of water data users, or for sub-communities with advanced needs.

Second, the governing body for the datacenter must both:

- Support the academic research community by providing services that cannot be reasonably undertaken in the context of academic research, and
- Avoid competing with academic researchers that engage in hydrologic research.

The line between "research" and "support" changes over time, and the governing body of the datacenter must adapt to these changes by either undertaking support for useful research projects whose development is no longer considered to be research, or even by ceasing involvement with a project that changes status from support to research. In the natural course of fulfilling its missions, the datacenter will provide the ability to pursue research ideas that arise from the community but are beyond the community's resources to pursue alone. The governing body of the datacenter must make a clear distinction between the datacenter's support for research activities and the research activities themselves. The latter are overseen by researchers outside the datacenter.

Relationship to Existing CUAHSI Efforts

Many attributes of the datacenter are already attributes of CUAHSI.

- Standards, software, and data sources are designed to empower science, by a community of scientists.
- Software is open source and freely available, and community contributions are both encouraged and fostered.
- The community gains a voice through CUAHSI in representing needs to external entities.

The datacenter represents a natural evolutionary step beyond CUAHSI as it operates now. So far, CUAHSI scientists have functioned as software developers with limited outside software engineering support. CUAHSI has attempted to "train scientists as software engineers"; many CUAHSI developers are "scientist-developers" who engage with both science and software engineering. This is problematic when one considers the value propositions for science and software engineering:

- For a scientist, the value is in developing publishable concepts.
- Software engineering is concerned instead with making cyberinfrastructure robust and reliable.

In other words, software engineering is not a publishable scientific goal. At this time, science and software engineering are in active conflict within the CUAHSI HIS project; CUAHSI scientists are engaging in all of the software engineering that they can afford in the context of being scientists, and even more is needed:

- Much CUAHSI software remains in prototype form. Documentation and testing are limited in scope. There is limited release engineering and pre-release testing.
- There is frequent loss of maintenance continuity, when student developers move on to other projects or graduate and undertake their own careers.

This is not a fault of the developers, but rather, a direct effect of the value proposition for participating.

As well, hydrology could benefit greatly from "data engineering," independent of the software engineering that provides robust technical infrastructure. Data engineering refers to those processes – automated or manual – that provide data to researchers in a useful form. Data engineering aspects include mechanisms for collection, quality control, and curation of data sources. These mechanisms are not solely technological, but also involve human expertise and effort.

Relationship to Other Data Management Efforts

While the datacenter seems to superficially solve the same problems as other data management efforts, including UniData[1], DataOne, and others, the datacenter faces unique problems that arise from the nature of the data being curated. Many of the data sources in HydroCatalog are best classified as "operational data sources".

1. There is often real-time or near-real-time reporting of measurements.
2. Quality control occurs after initial reporting.
3. Thus the contents of data sources change over time:
    a. By appending new data.
    b. By validating, correcting, and certifying older data.

Thus it is often the case that two queries from the same source for the same search parameters – separated in time – will contain different data.

This contrasts with the notion of a "data product" as defined in other data management efforts:

1. A data product is created only when data are corrected and validated.
2. Validation often involves cross-correlation with other external data sources.
3. Data products contain measurements for pre-specified intervals of time.
4. Data products do not change over time.

While there is community demand that the datacenter manage data products of this kind, this has not been the main thrust of CUAHSI efforts so far.

In cataloging and searching operational data sources, there are several challenges that make traditional "digital library" approaches to curation insufficient:

1. Data quality and validity can vary over time, even from a single source.
2. Data usage often involves fusing data from several sources into a coherent picture of a physical feature, including correcting for varying time scales, inferring values of missing data from other sources, and applying models to the result.

The result of this complex process is a "data product". Thus it might be appropriate to think of operational data as being "lower level" than a "data product."

A second important difference between CUAHSI HIS and other data management approaches in the Geosciences is that the data to be managed is not (yet) terascale or petascale in size. Consider, e.g., UCAR data and HydroNEXRAD[2,3,4]. One of the reasons that HydroNEXRAD remains difficult for users to apply is that it converts between terascale weather data and small-scale hydrology estimates. The datacenter may desire to take an active role in performing this kind of conversion for the community.

DESIGN AND STRATEGIC PLAN

The high-level design and strategic plan for the datacenter includes defining governance, processes, and features of the datacenter that satisfy the above requirements. In this section, we will define strategies; the next section will consider implementation and a tactical plan.

Governance

The governance of the datacenter will be patterned after that of CUAHSI. The ultimate authority for the datacenter will reside with the CUAHSI Board of Directors, as advised by its Standing Committees. The CUAHSI Executive Director oversees the datacenter director, who will in turn manage the above activities through hiring appropriate staff. It is important that the datacenter be associated with CUAHSI, and autonomous from all academic entities, to assure impartiality and that needs of diverse academic researchers are met without institutional bias.

The datacenter's mission will be a moving target, subject to periodic review. In the first years, the emphasis will be upon putting the ideas of the HIS project "into production", but the datacenter will simultaneously be looking out for new opportunities and directions that better serve its constituencies. The datacenter is – at its core – a member organization, and will evolve to serve the needs of its members.

CUAHSI governance is based upon meeting the needs of target constituencies and maintaining the value propositions of the datacenter, as documented below.

Business Model

The datacenter has several potential sources of funding, including:

1. Subscription services, including publication, usage tracking, and support.
2. Federal agencies seeking to empower researchers to do hydrologic science.
3. Federal contracts to build infrastructure to support national needs.

4. Strategic partnerships with industry centered upon development projects.
5. Contracts to produce specific software components for academia or industry.
6. Consulting services for end-users, developers, and data providers.
7. Fees for use of specific (advanced) services.

The datacenter benefits from being derived from two separate traditions: hydrologic science and open source software. Open-source software like CUAHSI HIS is typically funded not by selling the software, but instead by selling support, consulting, and customization services. For example, a scientist wishing support for developing a new plug-in could contract phone support for a small fee, and development help for a somewhat larger fee. Hydrologic science is often funded via strategic government investment in areas of national need. Thus there are several funding sources with good potential for funding the cyber-infrastructure activities that the datacenter is intended to undertake.

The proposed datacenter is subject to the following constraints:

- It must be sustainable in the long term.
- It must enable science without competing with scientific projects.
- It must be community governed and empower a constituent community.

Of these, the most restrictive constraint is that of sustainability. To be sustainable, the datacenter must balance the priority of its missions against potential for revenue. This – in turn – establishes the relative priority of processes within the datacenter, because some processes are more able to produce revenue than others.

Economic Drivers and Revenue Sources

Several required activities within a research project do not directly relate to research and scientific inquiry. The most important of these include:

- The need for a **data management plan** that makes data available to future research projects.
- The need to prove **broader impact** of scientific projects and discoveries.
- The need for **outreach activities** that explain scientific results to a broader audience.

So far, CUAHSI HIS has encouraged scientists to publish their own data and contributed software (HydroServer) for the purpose. In essence, this is equivalent to self-funding the data management plan from grant funds, which has several disadvantages:

- There is duplication of effort as many researchers have to engage in the same activity.
- The expertise to properly run a data publication source is not typically available in a hydrology research team.
- When funding ends, any data publication services are at risk of being terminated.

This provides an income opportunity for the datacenter:

- The datacenter will have **sufficient expertise** to provide comprehensive data management.
- Most hydrology researchers need more or less the **same data management services,** which lead to eventual publication of data for public use.

- Data management is subject to **economy of scale**; managing data for multiple projects is not substantively more difficult than managing data for one project.
- The datacenter – unlike academic laboratories – can by its nature guarantee that published data will be **available in perpetuity**, past the end of academic research projects.
- Thus, the datacenter can provide comprehensive data management at a **fraction of the cost** of self-managing and self-publishing data.

HydroServer is useful as a reference standard but – in practice – it is too much work to deploy for small-scale hydrology projects. There is also a constant burden after deployment of updating the service to address bugs and changes in standards. Medium-scale projects that can afford to deploy HydroServer do not necessarily make sufficient provision for making data available in perpetuity. Evidence for this comes from the fact that data sources regularly disappear from the HydroCatalog and the catalog must be edited. Thus we conclude that HydroServer is not a sufficient data publication mechanism for small-scale or medium-scale data collection efforts. Further, there is strong evidence that a very large amount of data gathered by smaller research projects – that are independent of the national observatories – is never published in a usable form. Thus:

- Data management and publication should be made available at the datacenter on a subscription basis.
- The datacenter should maintain mechanisms for easy transition from publishing data at remote HydroServer installations (or via other compatible mechanisms) to publishing data in cloud-based data storage managed by the datacenter.
- The widespread adoption of HydroServer should be deprecated, in favor of centralized data management, which has the advantages of outliving specific research projects and providing a high level of data management with high economy of scale.

Once the above decisions are made, related revenue sources become available as a result. For example, the second need on the above list is the ability to prove impact. The current CUAHSI HIS catalog keeps usage statistics on the HydroCatalog and which data are downloaded by which entities. This is valuable information for research projects because data use is direct evidence of broader impact. Thus:

- The datacenter should provide subscription services that provide high-quality usage tracking data for the data that it curates.
- This should include tracking use of wrappers and other non-standard data access methods.

There is also a need for some simple help for research laboratories that are already running their own data publication software. The evidence from HydroCatalog is that backup and restore for data services are problematic. Quite frequently a data source listed in HydroCatalog will be offline for an extended period. This provides a revenue opportunity:

- The datacenter should provide subscription backup services for data currently provided via WaterOneFlow services and other formats.
- The datacenter should provide a subscription service for automatic failover capability to backup data inside the datacenter cloud, should remote publication services go offline.
- The datacenter should provide subscription services to maintain wrappers that translate from non-standard publication formats to WaterOneFlow services.

- There should be a seamless transition plan from hosting data remotely to hosting it under control of the datacenter.

Finally, outreach is another key component of a successful research project. The datacenter can help in this regard by:

- Developing and maintaining high-accessibility data discovery clients for use by general audiences.
- Making materials and data sources available to introductory classes in Environmental Science and related disciplines.

The above considerations drive one more policy decision which is at its core a cost-saving measure. One disadvantage of related efforts at centralized data management is to attempt to implement relatively complex data ownership policies for pre-publication data. The only way that this datacenter can achieve economy of scale is to sidestep such complexities. Thus the data management policies of the datacenter center upon eventual public availability of data, with limited privacy options in support of eventual academic publication:

- All data managed by the datacenter via subscription will be made available to the public after a predetermined time.
- Data will automatically be published at the end of any subscription.
- The datacenter will not manage data that might not ever be made public.

Storing data invokes a cost, while publishing it provides community value. The datacenter should not involve itself in data management efforts that do not provide that value. It is our estimation that providing data management services for perpetually private data is not cost-effective, due to the complexities involved and the relatively low value to the datacenter's constituencies.

Subscription-Based Revenue

The needs of the community suggest a subscription-based business model that exploits eventual economy of scale for service provision after a brief transition period. The business model for the datacenter is built upon the following claims:

1. Continuity and robustness of data access has tangible financial value to scientists and can be sold as a subscription service.
2. Reporting of data access statistics has tangible financial value to scientists and can be sold as a subscription service.
3. Both of these services scale well with respect to human resources within the datacenter.
4. The appropriate fees for services should remain small, in hopes of capturing the whole market and thus exploiting that economy of scale from a human standpoint.
5. Remote data publication for small-scale and medium-scale projects should be deprecated in favor of datacenter-based data publication.
6. To keep fees low, data management policies should be centered around eventual data publication, to avoid complexity and better serve the community.
7. Non-subscription services should be funded from the income from subscription services.
8. Costs of subscriptions should be keyed to the cost of storage, the cost of human labor necessary to maintain the terms of the subscription, and an additional component to fund non-subscription services that are made available at the datacenter without charge.

The hope is that the cost of subscribing to a datacenter service would be 1/3 to 1/10 the cost of providing the service in house, while at the same time providing the service at a greater level of security, robustness, and continuity than would be achievable in-house by a typical hydrology research laboratory or project.

The main datacenter revenue is from data hosting and reporting for small-scale projects, not from software development. Software maintenance is funded in part from subscriptions, and thus takes a secondary role to that of providing and managing data, and providing comprehensive catalogs and data discovery services.

Strategic Partnerships

A final facet of the datacenter's business model is the formation of mutually beneficial strategic partnerships with industry and research partners, to:

1. Prepare research developments for general use by the hydrology community, including innovations in:
    a. Data discovery and analysis software.
    b. Data product creation and validation.
2. Make business products more usable by the academic research community, including:
    a. Measurement storage, publication, and sharing for vendor-specific devices.
    b. Integrating datacenter innovations into vendor data analysis products.
    c. Adapting the data products of major environmental observatories to be discoverable and usable by academic researchers.

These will ideally be true partnerships, with both the datacenter and the researcher or vendor providing some of the resources necessary for implementation. The datacenter will provide the support and continuity aspects, while researchers and vendors will provide the scientific perspectives necessary to transform knowledge into services.

PROCESSES

In the following sections, we will discuss strategies and tactics for fulfilling the missions of the datacenter as defined previously. In each case, "strategies" are long-term decisions that are not likely to change over the lifetime of the datacenter, while "tactics" are short term decisions that are subject to change and periodic review.

Processes are those tasks that are assured by human staff in the datacenter, with some assistance from computing infrastructure. Strategic analysis of these processes includes defining human roles that aid in assuring that the process operates smoothly. Roles are not job descriptions – but rather – parts of job descriptions. In the following, we will decompose the business processes of the datacenter into roles, policies, and tasks required to enable a role. In the following, the word "define" entails making a policy decision, while "create" means to produce a tangible, non-policy output.

At the highest level, the datacenter will accomplish the above missions by engaging in the following somewhat distinct business processes:

1. **Standards curation:** the process of proposing, documenting, and supporting standards and best practices for data publication and data access that empower the academic research communities in hydrology and related disciplines.

2. **Data source curation:** the process of ensuring that data sources represent accurate, timely, and useful information, including standards-compliant wrappers for non-standard data services.
3. **Data product curation:** the process of ensuring that contributed data products of research projects continue to be accurate and available to users in perpetuity.
4. **Data catalog curation:** the process of ensuring that data catalogs contain accurate, timely, and useful information on available services.
5. **Service administration:** the process of ensuring that data services provided by the datacenter remain reliable, robust, and available to researchers.
6. **Software maintenance:** the process of ensuring that software remains reliable, robust, and usable in the presence of updates and new features.
7. **Community support:** the process of helping users effectively utilize the software, and helping developers to implement new approaches to data publication, sharing, discovery, and analysis.
8. **Outreach and advocacy:** the process of ensuring that potential users are made aware of the available water data resources, and that data providers are made aware of issues that affect academic use of their data.
9. **Contract management:** the process of negotiating appropriate contracts between the datacenter and outside providers, including both services contracted by the datacenter and services provided by the datacenter on a contract basis.

Many of these processes are already undertaken by CUAHSI member-scientists on a volunteer, "best effort" basis. The datacenter will assure these processes as its core mission.

In the following, we will discuss each process in detail, including both strategic and tactical analyses. The strategic analysis for a process includes considerations that are not likely to change, while its tactical analysis represents a first cut at a set of governance decisions that should be made in order to implement strategies.

Standards Curation

Strategic overview

Scope: The datacenter will support the ongoing effort of CUAHSI to develop standards as a research community, but will defer to CUAHSI for governance decisions. The datacenter's support for standards will include documenting and disseminating existing standards, technically evaluating new standards and changes in standards, and testing standards in practice via changes in software and services. These efforts will address standards for

- Data publication (e.g., WaterOneFlow services).
- Data format (e.g., WaterML).
- Data search and discovery services (e.g., as embodied in HydroCatalog).

Status: Many members of CUAHSI are involved in standards development. The Geospatial Computing Group at the San Diego Supercomputing Center (SDSC) has played a central role, partly because of their involvement in providing standardized services, which gives them a unique understanding of the technical impact of a standard. CUAHSI members maintain a presence in all standards efforts, including definition of WaterML2.0, the Open Geospatial Consortium (OGC), and others.

Strengths: It can be argued that one of the central contributions of CUAHSI to the hydrological sciences is its role in proposing, evaluating, and implementing standards to enable data sharing.

Critique: Standards activities are a secondary responsibility for all of those involved. Thus, involvement in standards now occurs on a "best effort" basis.

Strategic recommendations

The standards activities of CUAHSI are central to the value of CUAHSI and the eventual value of the datacenter. So far, CUAHSI has concentrated on building software that embodies standards, and there is some sentiment that the standards are a more important contribution than the software. Thus, the datacenter will document, evaluate, and curate standards separate from the software that embodies these standards. This requires the following tasks:

- Define the datacenter role of **standards curator** (separate from the task of maintaining software and embodiments of standards).
- Define workflows for standards curation, including documentation, publication, and periodic review.

Tactical overview

This is an area in which new standards need to be developed, notably:

- Standards for describing data quality.
- Harmonization of hydrological standards with other geospatial standards.

Nonetheless, this is an ongoing process with a non-varying time commitment. It is front-facing, in the sense that there is a need for ongoing discussion with other standards organizations and major data providers. No transition planning is required.

Tactical recommendations

- **Subcontract standards curation** to the Geospatial Computing Group at SDSC. This group has proven effective in creating standards and in curating reference implementations of those standards.
- Interface with SDSC via a data source curator (described below).

Economic prognosis

- Cost: ½ FTE within datacenter for liaison duties, SDSC subcontract, travel.
- Revenue: indirect, from subscribers to other services that depend upon the standards activity.

Data Source Curation

Strategic overview

Scope: Data source curation is the process of assuring that data sources provided by the datacenter remain functional and available in the presence of changes. The scope of this curation activity includes curating and maintaining:

- "Operational" research data sources, stored as a service of the datacenter.
- "Legacy" operational data sources for servers and services that no longer operate (or become temporarily unavailable).

- Wrappers of current data sources that provide data from non-standard sources in standard formats.

There is a fundamental distinction between data source curation (which concerns what might be called "operational data sources" that change over time) and data product curation (which concerns datasets that have been compiled as output from research projects). For more details on the latter, see the discussion below.

Status: Data source curation is currently performed by the SDSC Geospatial Computing group. There is currently no support for internal operational data storage, backup and restore, though there is limited support for storage of legacy data upon request.

Strengths: It can be argued that the success of CUAHSI HIS as a whole depends upon providing paths to standardized data publication other than utilization of the standardized data publication server HydroServer. This includes writing and maintaining wrapper services that standardize the output of non-standard services, as well as customizations of HydroServer that run externally.

Critique: Maintenance of wrappers and other adaptations is done on a best-effort basis and based solely upon user reports. While there has been timely response to user reports of wrapper malfunction, there is not a proactive, regular process for review of whether wrappers remain functional or remain semantically correct in the presence of unpredictable and uncoordinated changes in the data sources.

Strategic recommendations
The efficacy and usefulness of data sources is mission-critical to CUAHSI and the datacenter. Thus, assuring this crucial service should be a separate role.

- Define the role of **data source curator**, with the responsibility for the proper function of data sources maintained within the boundaries of the datacenter.
- Define the role of **wrapper curator**, with the responsibility for the proper function of transformative wrappers crafted by the datacenter and its affiliates.
- Define **workflows for source and wrapper curation** that regularly checks sources and wrappers for correct function and mapping.
- Create **administrative tools for wrapper curation** for use by data source curators, including the abilities to:
    - View the semantic mapping for a wrapper.
    - Check the results of a wrapper via before-and-after views.
    - Refer problems with wrappers to the appropriate engineer within the datacenter.
- Define a **business model** and contractual details for internal storage of operational data for active research projects. This business model may include a model for contracting the datacenter to provide new wrappers of useful data sources that are not currently available in a standardized form.
- Partner with instrumentation and sensor vendors to create custom services targeted at storing and curating data from a variety of vendor data collection devices; make these services available via contract to individual researchers and institutions.

Tactical overview

According to the mission documented above, there are three parts to data source curation:

1. Operational data hosted by the datacenter.
2. Operational data from legacy sources.
3. Wrappers that expose standards-compliant interfaces to non-standard sources.

As well, for tactical reasons, there are two more parts:

4. Backup, failover, and restore services for existing data sources.
5. Seamless data publication for vendor-supplied data measurement units.

There are several policy questions that affect the cost of this process:

1. How much operational data will be hosted? How will the datacenter decide what to host?
2. What provision will there be for keeping operational data private at the wishes of researchers?
3. What will be the business model for hosting?
4. How will legacy data be selected for hosting?
5. How will wrappers be managed and curated?

The tactical plan for the datacenter is based upon a subscription-based business strategy that:

1. Gracefully degrades provided services when a subscription expires.
2. Provides value to the community even for expired subscriptions.
3. Provides strong incentives to re-subscribe when feasible.

The subscriptions for data publishing are based upon three tenets:

1. There is academic value in understanding use of published datasets.
2. There is no increase in cost in tracking use even if tracking is not paid for.
3. Thus, ongoing tracking can be accessed via a future subscription.

In this context, we make the following tactical recommendations.

Operational data policies

Whether operational data will be hosted will be a governance decision with a formal application process. There is a mission advantage to the datacenter to host as much data as it can afford to host. Therefore, hosting of data will be a subscription service with a low yearly subscription fee, to encourage broad adoption by small-scale research projects. The hosting service will be funded long-term by economies of scale: more subscribers do not require more human work from the hosting service. For subscribers, the datacenter will:

1. Provide a secure interface for uploading and revising data.
2. Guarantee to host and publish data in perpetuity.
3. Keep data private to the subscriber until a date specified by the subscriber.
4. Publish data according to the most current data access standards, updated as necessary.
5. Ensure that catalog entries remain up to date for the data.
6. Provide access reports for published data to the subscriber after the data becomes public.

If a subscription lapses, the data will still be hosted, but

1. Data will become public upon end of subscription.
2. Updates to the data will not be allowed without renewal.
3. Access reports will not be provided without renewal.

Hosting data after end of subscription is semantically equivalent to hosting legacy data.

Legacy data policies
The datacenter will host legacy data on a per-request basis. The ideal request for legacy storage occurs before the service is discontinued; then the datacenter can simply query the service and store the results. Post-discontinuation requests are more costly to handle and will require governance decisions. Original data owners can subscribe to data reporting services for a small yearly fee. For approved subscribers, the datacenter will:

1. Download and store legacy data in standard form.
2. Host the data in perpetuity.
3. Provide the data according to current data access standards.
4. Track access to data and – for subscribers – report access details to the original data owner.

If a subscription lapses, then:

1. Data will continue to be available.
2. Accesses will continue to be tracked.
3. Access reports will not be provided to former subscribers.

Access reports can be obtained by renewing a subscription.

Wrapper curation policies
Deciding what wrappers to maintain will be a governance decision. There are two ways that a wrapper can come into existence:

- By community demand, in which case it is provided free of charge.
- By request of a data provider, in which case the wrapper is provided as a subscription service for a small yearly subscription fee.

For subscribers, the datacenter will:

1. Create wrappers for specified data services.
2. Maintain wrappers over time to comply with current data access standards.
3. Ensure that catalog entries remain up to date for the wrapper.
4. Monitor access to wrapper services.
5. Report data access results to the subscriber.

For community-requested wrappers, the datacenter will still monitor access to wrapper services, and that data will be available to data producers for a subscription fee.

Backup and restore services
The original aim of CUAHSI HIS was to encourage researchers to run their own data source servers. The recent advent of cloud technologies forces us to reconsider that aim. In fact, hydrologists and even major labs

lack the resources to provide a persistent and sufficiently robust data source. Thus, the datacenter can add value to HydroServer by:

1. Providing backup and restore services for HydroServer for a yearly subscription.
2. Implementing data storage in the cloud.
3. Providing failover data access in case of original server failure, to avoid loss of continuity.
4. Providing legacy data access for retired servers once a project is over.

The business justification for this is that:

- Many HydroServer sites actually spend 1/3 of what they would need to spend for high availability and persistence.
- Due to economy of scale, the datacenter can provide that service at less than 1/10 of the cost for an equivalent quality of service at an individual site.

Thus, for business reasons:

- The price for backup and restore services should be kept to less than 1/10 of the price of accomplishing the same quality of service in isolation.
- The datacenter should actively encourage sites to upgrade to hosting in the datacenter's cloud, with a similarly aggressive financial incentive.
- The datacenter should offer conduit-writing and maintenance subscriptions for maintaining data conduits from sites that do not utilize HydroServer.

Direct-from-device publication services

While most vendors provide data recording services for their devices, these solutions stop short of making data seamlessly available in a data search framework. The datacenter can provide value by creating custom data publication systems that retrieve data from specific vendor devices. Such operational data can be kept private to a group of researchers for a specified time, after which it becomes public data. This service can be offered to users of the major vendors of hydrology-related and meterology-related sensing equipment, handling the most common equipment types first. The vendors will have a strong incentive to cooperate, given that their instruments will then enjoy an easy path to adoption.

Tactical recommendations

- Assign one in-house curator to oversee all five processes.
- Utilize datacenter programming resources to create:
    - Compelling and useful "dashboard" displays for each data source that document use of data sources.
    - Online backup and restore services for HydroServer.
    - Conduits to non-HydroServer data sources, by subscription.
    - Curation interfaces for viewing the status of data sources.
    - Custom publication subsystems that publish data for specific kinds of data collection equipment.
- Subcontract wrapper maintenance and documentation to the Geospatial Computing Group at SDSC.

Economic prognosis

- Cost: one in-house FTE, subcontract to SDSC for one-half FTE for wrapper maintenance.
- Revenue: direct, through subscriptions.
- Funding source: direct, through subscriptions, as well as subsidies for the backup service, to encourage its use and encourage smooth transition to cloud-hosting for legacy data.

Data Product Curation

Strategic overview

Scope: Data product curation is the task of curating and making available derived data products that are the result of research projects and related activities. Unlike data sources, which involve potentially real-time data collection and dynamic updates, most data products are – at the current time – files of data that have been collected, studied, modeled, and frozen in a specific configuration for further study. Thus, while data sources are dynamic entities in the external world, data products are documents stored at the datacenter that represent the state of knowledge at a specific time. Current work in sharing models of geophysical systems may provide another form of data product that is a reusable model connected to appropriate data sources.

Status: no data product curation is provided within CUAHSI at the present time.

Strategic recommendations

- Define the role of **data product curator** responsible for ensuring that data products are appropriately preserved and remain available in the future.
- Define a **data product workflow** for ensuring availability and usefulness of the data products stored inside the datacenter.
- Strategically determine data products that should be **created and maintained by the datacenter** for community benefit.
- Define a **business model** for data product storage based upon perceived value and cost of perpetual storage.
- **Leverage existing infrastructure** for data product curation, including the DataOne services for distributed availability of data products and research results.

Tactical overview

Data product curation will be a free service of the datacenter, funded indirectly through other revenue streams. The most straightforward way to support data product curation is for the datacenter to administer a DataOne node that can be linked to its operational data sources for reference purposes. The curation will be self-service; members can upload their papers and data products themselves. Because this is a self-service curation, very little staff time will be required in the beginning.

However, the complexity of creating certain data products – e.g., by interaction with the HydroNEXRAD system for extracting water data from weather data – suggests that the datacenter should have a role in creating data products of general interest. Creating and maintaining these data products is a target goal for years 4-5 of the datacenter. This will require strategic partnerships with researchers whose role is to define and refine the products.

Economic prognosis

- Cost: initially minimal, due to self-service model for uploads. In final operation, ongoing data product creation will be a fixed cost within the datacenter's operations.
- Revenue: indirect, by encouraging users to utilize other subscription services. It is possible that the datacenter will enter into direct agreements to produce certain data products on contract.

Data Catalog Curation

Strategic overview

Scope: The existing CUAHSI HIS contains an extensive catalog of data sources that requires ongoing curation, validation, and updates. This includes data sources that are maintained by the datacenter itself, that are stored within the datacenter or implemented as standards-compliant wrappers for non-compliant sources. The datacenter will periodically check the existing catalog for accuracy, and maintain the software that periodically updates the catalog for changes in sources, as well as the software that provides standards-compliant wrappers.

Status: Catalog curation is currently done by the Geospatial group at the San Diego Supercomputing Center. This includes checking that data services are running, harvesting metadata from data sources, and reporting on overall data usage.

Strengths: HydroCatalog is designed appropriately for the addition of a curation component. The software implementation remains relatively straightforward and simple, and it will not be difficult to add a curation support subsystem to it.

Critique: Curation is done by SDSC software specialists, using primitive back-end software to correct database problems. Problems have often been addressed by modifying the databases using structured query language. This has been done reactively, responding to issues discovered by users and developers. There is currently no interface whereby a hydrologist outside SDSC could participate in curation of data. The exact workflow for curation, including assurance that HydroCatalog remains current in the presence of external changes in services, is not clearly defined, nor is there a user interface to support that workflow. There is no process for periodic review of wrappers for correctness. There is no workflow to regularly assess the function of data source wrappers. The result is that users lose time by interacting with long-inactive data sources during data discovery.

Strategic recommendations

Addressing these concerns requires the following tasks:

- Define the datacenter role of **data catalog curator,** who assures that HydroCatalog remains useful in the presence of changes in the behavior of external data providers.
- Define a **workflow for catalog curation** that includes policies for removing non-functional HydroServers from HydroCatalog, and generally insuring standards for data quality and availability.
- Create **administrative tools for catalog curation** for use by catalog curators in accomplishing the workflow, including the abilities to:
    - Test the state of a data source and determine its appropriateness for inclusion in the catalog.

    - Set the state of a data source, and thus affect how it is listed in the catalog.
- View and modify the mapping between a wrapped service's input and output.

Tactical overview
Data catalog curation is a high-cost activity that is indirectly funded from other datacenter revenue. It provides value to subscription services for data curation, by insuring that curated data will be found when desired by researchers. Thus there is high motivation to maintain a very high-quality catalog that is up to date and thus provides that value.

Note that catalog curation is an indirect revenue center, because the act of curation discovers inactive data sources that could be hosted and monitored through subscription.

Economic prognosis:
- Cost: ideally 1-2 FTE, domain specialists in hydrology.
- Revenue: indirect, through locating potential subscribers via the act of curating the catalog.
- Source of funding: indirect, from other subscriptions.

Service Administration

Strategic overview
Status: Another part of the current CUAHSI HIS is a set of running services that provide search capabilities for the CUAHSI HydroCatalog. These services must be maintained in a state of readiness, and must be tested after each change in the software.

Status: HydroCatalog is currently hosted by the Geospatial Computing Group at the San Diego Supercomputing Center. This team is responsible for HydroCatalog and the harvesting software that populates the catalog. Sites are registered in HydroCatalog via a simple front-end program. Every week, the catalog is updated by re-harvesting metadata from each functional HydroServer instance.

Strengths: HydroCatalog software is relatively cohesive (in the sense that it accomplishes one task, and accomplishes it well). It is well-designed and its function is straightforward to test. Maintenance interfaces have already been created that report on the health of provided services.

Critique: Hosting of the CUAHSI HIS services is not an ideal role for the SDSC group. They are better utilized in upon defining appropriate data standards, as well as bringing supercomputing to bear on some of the grand challenge problems that face hydrology.

Strategic recommendations
The roles of developer, service administrator, and catalog curator should be separated, and appropriate interfaces provided for each role.

- The **software developer** recommends changes in software, based upon the needs of the datacenter's constituencies as specified by the CUAHSI board.
- The **service administrator** manages availability of the service, and recommends hosting options that maximize that availability at minimal cost.

- The **catalog curator** (defined previously) ensures that data in the HydroCatalog is up to date, points to functional HydroServers, and that servers are correctly categorized for use in data discovery.

It is possible that in practice, one person may take more than one of these roles, but the point of this recommendation is that that coupling is not strategically essential.

This includes the following tasks:

1. Define the role of **service administrator**, as well as workflows for service administration that include:
    a. Defining service objectives for behavior of services.
    b. Monitoring services to assure that objectives are met.
    c. Monitoring service usage and creating usage reports.
    d. Coordinating service changes to avoid disruptions.
2. Host the services in a true datacenter environment, rather than a prototyping environment.
3. Create interfaces that report whether objectives are being met, and create usage reports.
4. Coordinate changes in HydroCatalog and the other components, including centralized testing of interactions between HydroCatalog and the other components.

Tactical recommendations
The tactical plan for service administration is to:

- Subcontract data hosting and search services to a major cloud storage provider.
- Maintain in-house staff to monitor and update the services.

Economic prognosis

- Cost: ½ FTE, long term, plus cost of subcontract, which will be based upon amount of storage.
- Revenue: indirect, from subscriptions.
- Funding source: indirect, from subscriptions.

Software Maintenance

Strategic overview
Scope: The existing CUAHSI HIS includes a software suite for cataloging, searching, and accessing water time series data. CUAHSI HIS is currently:

- An embodiment of current "best practices" for data publication, discovery, and/or use.
- A testbed for new techniques for data publication, discovery, and/or use.
- A focus for community contribution and sharing of best practices.

CUAHSI HIS software – like any other software – requires periodic maintenance. Maintenance includes repairing problems, responding to user requests for improvements, and incorporating changes from the community.

One goal of the datacenter is to maintain and support the CUAHSI suite of software tools. This suite is comprised of three main parts:

- **HydroCatalog:** a centralized, standards-compliant data service that stores the Internet locations of water time series data.
- **HydroServer:** a sufficient implementation of a standards-compliant water data source that can be listed in HydroCatalog.
- **HydroDesktop:** a data discovery and modeling client that enables discovery of resources recorded in HydroCatalog and (typically) provided via HydroServer.

While HydroServer and HydroCatalog implement relatively straightforward sets of services, HydroDesktop includes both an embedded GIS subsystem and a complex "plug-in" architecture that allows users to become developers and write their own "plug-ins" to extend its function, much in the manner of a web browser such as Firefox. This architecture has empowered hydrologists to extend the desktop in ways that would be impractical without the HydroDesktop framework.

Status: The three different software systems comprising CUAHSI HIS are currently developed and maintained by three geographically-distributed entities:

- **HydroCatalog** is maintained by the Geospatial team at the San Diego Supercomputing Center (SDSC).
- **HydroServer** is maintained by Jeffrey Horsburgh at Utah State University.
- **HydroDesktop** is maintained by Dan Ames at Idaho State University.

Strengths: Geographically-distributed development has encouraged the teams to minimize coupling between HydroCatalog, HydroServer, and HydroDesktop, thus conforming to appropriate software engineering practice. During prototyping, the groups have been able to function more or less autonomously with minimal coordination.

Critique: While appropriate to a prototyping phase, distributed governance of this kind threatens to impede the progression from prototype to product.

- The use of student labor has led to loss of institutional memory regarding a component or components. When a student leaves, a component no longer has a maintainer.
- There is no academic value proposition in the labor-intensive process of assuring software quality; the result is not "publishable".
- Distributed maintenance has proven problematic when changes need to be coordinated. E.g., a change in HydroServer must be coordinated with the mechanism in HydroCatalog that harvests it, and the interface in HydroDesktop that recovers the data. Thus there is a need for centralized governance of testing and software releases.

There is a problem with testability of HydroDesktop that suggests design changes to support testability. In particular, the ideal plug-in architecture for HydroDesktop is a pure "star" configuration in which each plug-in can be present or not (as in browsers such as Firefox). This is not the case now: some plug-ins depend upon others and some are so crucial to overall function that their installation is no longer optional. This affects the testability of HydroDesktop and makes it difficult to properly enumerate and test all reasonable usage scenerios.

There is also a problem with the way HydroCatalog metadata is defined, that centers around the identity of a data series. When a data source agency publishes a data series in HydroServer, the sole identity of that data series is the metadata that the agency provides in the HydroServer configuration. If that agency then changes that configuration, the data series can seem to disappear from HydroDesktop, while a new series seems to appear. There is no effective mechanism for maintaining the identity of a data series in the presence of these changes. Addressing this problem will require coordinated changes in both HydroServer and publication standards, as well as in HydroCatalog.

Another deep question concerns the role of HydroDesktop as an embodiment of search standards and best practices. It is likely that a simpler embodiment is needed. HydroDesktop is made more complex by being built upon the powerful "dotSpatial" GIS system. This makes it more useful to "power users" who are familiar with GIS, but less accessible to users who simply want access to water data. Because of the intimate relationships between HydroDesktop and "dotSpatial",

- HydroDesktop requires a substantive investment in time to learn.
- It is relatively difficult for a new developer to extend it (though there is high potential value in doing so).
- The standards being embodied are obscured by the complexity of the embodiment.
- It is difficult to port to platforms other than Windows and Microsoft .Net.

Thus, HydroDesktop is not an ideal embodiment of search and discovery standards for a large segment of users.

Strategic recommendations

The distributed model of governance for CUAHSI HIS should be replaced with a more centralized governance model based upon industry best practices for software quality assurance. The teams that currently engage in distributed development will continue to do so, but their work will be considered prototypical until re-engineered for production by the datacenter's engineering team. The datacenter's team will include some new roles:

1. One or more **project librarians** will be responsible for documenting, building, testing, and releasing overall builds of the software. This librarian will function according to policies for documenting, building, testing, and releasing versions of the software. This includes:
    a. Coordinating software changes between interacting subsystems.
    b. Centralized testing and release management of the suite as a whole.
    c. Use of service versioning to avoid version skews between versions of components.
    d. Bug and issue tracking and documentation.
2. A team of **software developers** ("software engineers") will
    a. Oversee the adaptation of HIS for better testability.
    b. Implement new components to aid in curation and testing.
    c. Document and maintain institutional memory for incorporated software.
3. One or more **developer support engineers** will
    a. Assist scientist-developers with constructing prototypes, including person-to-person support, as well as help with testing, release planning, and integration into the overall suite of software.
    b. Re-engineer prototypes for production use, with or without collaboration with the original developer.

These roles have perhaps different lifespans. The librarian role responds to changes proposed by the community and is essentially permanent, for the lifetime of the DataCenter. The re-engineering required to turn CUAHSI HIS into a software product is more limited in scope, and can be accomplished in 1-2 years by the appropriate professionals. Support for developers and re-engineering of prototypes is an ongoing process for the lifetime of the datacenter.

Tasks involved in assuring this process include:

1. Define the role of **project librarian**, including defining workflows for testing, release management, and release scheduling.
2. Define the role of **developer support engineer**, including workflows for developer support and a careful definition of the scope and limits of that support.
3. Define the role of **software developer**, including limiting the contact of core developers with outside developers in a support role.
4. Make a **governance decision** for the single-platform Windows search client HydroDesktop about further support and development. At this point, it seems that the datacenter should support – but not extend or re-engineer – the existing HydroDesktop; that seems to be a research goal outside the main missions of the Datacenter. Further development of HydroDesktop should instead be the result of an ongoing partnership between the Datacenter and the current lead developers on the project, who remain highly motivated to continue development.
5. Define and develop a **multi-platform version of HydroServer** to make publishing data more accessible to researchers who do not possess Windows server infrastructure.
6. Define and develop a **multi-platform data discovery client** that functions in all modern desktop and laptop environments: Windows, Linux, Mac, and potentially also Android, IOS (IPad), and Chrome. This may be web-based.

Tactical overview

Software maintenance is one of the more controversial aspects of the center, because of an active tension in CUAHSI HIS development between:

- Highest common denominator search clients like HydroDesktop.
- User requests for lowest common denominator search clients like HydroSeek.

In the tactical plan, the cost/value tradeoffs determine the datacenter's tactical direction:

- Subscriptions will be purchased only if hosted data is used.
- Data will be used if there is an accessible client for accessing it.
- Thus, the economic health of the datacenter depends upon supporting a highly accessible client.

Tactical recommendations

- Continued maintenance of server software including HydroServer and HydroCatalog.
- Development and support of a highly accessible, multi-platform web based client.
- Partner with researchers to support highest common denominator clients such as HydroDesktop.

Economic prognosis

- Cost: Ideally 2-3 FTE programming staff, subcontracts to SDSC, Idaho State, and Utah State for ongoing consulting services.
- Revenue: indirect, through subscriptions to data services and/or support services.
- Funding source: indirect, including transition funds from NSF and ongoing funds from subscription services.

Community Support

Strategic overview

Scope: Community support entails assisting CUAHSI members in utilizing the services of the datacenter, including standards, software, and data. Services range from training for novices, to advanced assistance in publishing data and/or writing data search tools. Statistically, few sites utilize generic CUAHSI HIS services to publish their data; instead, they modify the provided software for site preferences.

Status: Community support is currently provided both by staff of CUAHSI and CUAHSI members, on a best-effort basis. User forums are available on the CUAHSI website.

Strengths: Support has enabled many users to begin to utilize CUAHSI software, and has resulted in many success stories.

Critique: Support is provided on a best-effort basis by principals who have other primary roles. There is no person truly "responsible" for user support.

Strategic recommendations

Assuring user support requires the following tasks:

- Define the datacenter role of **user support engineer.**
- Define workflows for user support, including request ticketing, guidelines for general support accessibility, scope of available support, and a triage strategy for escalating requests to other roles including developer support engineer, service administrator, and software developer.

Note that user support is a separate task from developer support as defined above.

Tactical overview

There are several different modes of user support:

- Free, "best effort" support from datacenter staff.
- Free, "forum" support from other users.
- Support subscriptions for user and/or developer support.
- Subcontracts for tools that locate data with specific attributes.
- Strategic partnerships with industry on projects of mutual benefit.

Tactical recommendations:

- Combine roles of user and developer support for free ("best effort" services).

- Hire or subcontract additional support staff in accordance with subscriptions, partnerships, and subcontracts.

Economic prognosis:

- Cost: moderate (1 FTE) for free services, self-funding for subscriptions.
- Revenue: direct, from subscriptions and subcontracts.
- Funding source: partly direct, partly indirect through related subscriptions.

Outreach and Advocacy

Strategic overview

Scope: Outreach and advocacy entail making researchers aware of the capabilities and potentials of the datacenter to aid in their research. This includes advertising the availability of data services, holding public tutorials and training, and maintaining a presence in the research communities that are potential clients of the datacenter.

Status: Outreach and advocacy is currently undertaken by CUAHSI and member scientists on a best-effort basis.

Strengths: The outreach program includes regular webinars on use of search and discovery tools, as well as an established presence at relevant conferences in hydrology and related disciplines.

Critique: There is no overall strategy for outreach; it is done only when convenient. Presence at events depends entirely on whether a principal in the CUAHSI projects can attend.

Strategic recommendations

- Define the role of **outreach coordinator,** as well as workflows for defining outreach strategy, implementation, and evaluation.

Tactical overview

In a datacenter funded by subscriptions and contracts, outreach is crucial. The outreach coordinator must:

- Identify and target potential markets.
- Formulate targeted advertising.
- Make first contact for strategic partnerships with industry.

Tactical recommendations

- Retention of one FTE with hydrology knowledge and public relations experience.
- Formation of an outreach plan, in coordination with the contract manager (described below).
- Institute data collection on user attitudes toward subscription services.

Economic prognosis

- **Cost:** 1 FTE plus travel.
- **Revenue:** indirect, from subscriptions and subcontracts.
- Funding source: indirect.

Contract Management

Strategic overview
Scope: Contract management is the task of managing contractual obligations between the datacenter and its subcontractors (e.g., to provide services) and between the datacenter and users (e.g., to provide support).

Status: There is no analog to contract management of this kind in CUAHSI now, although it does have a business manager. The kinds of contracts that the datacenter will need to utilize and assure are not yet defined or in use.

Strategic recommendations

- Define the role of **contract manager** who will oversee and be a signatory authority for the contracts required for the datacenter to operate, as well as business processes for periodic contract review and renewal.

Tactical overview
The current tactical plan is based upon several kinds of legal agreements, including:

- Data publication, wrapping, archiving, and backup subscriptions.
- Support subscriptions.
- Strategic partnerships with industry.

Thus, the datacenter will be managing a large number of relatively small contractual obligations. The contracts should be simple to understand, and yet clearly specify the terms, liabilities, and consequences of non-compliance.

Tactical recommendations

- Retain one FTE with legal/contract experience (It is likely that the CUAHSI Business Manager will fulfill this role through a subcontract relationship).
- Craft contracts according to tactical considerations listed above.
- Institute a Customer Relationship Management plan for maintaining business contact, including newsletters, updates, and other regular communications.

Economic prognosis

- **Cost:** 1 FTE with legal expertise, front-facing.
- **Revenue:** indirect, through negotiation of subscriptions and subcontracts

TACTICAL OVERVIEW

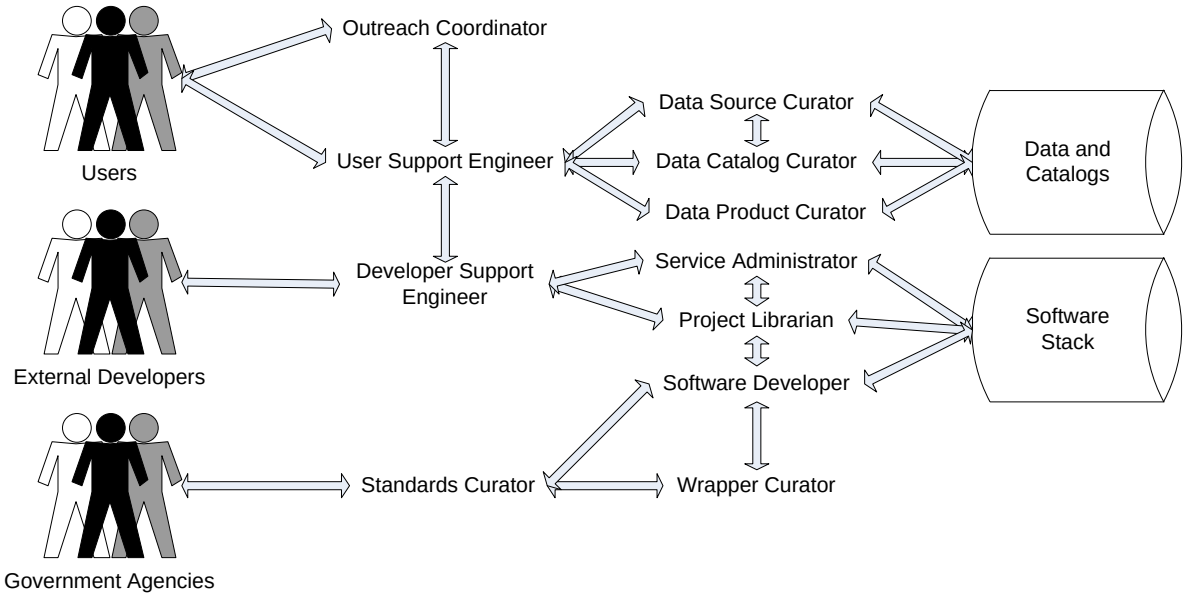
Roles

The above roles are not job descriptions – but rather – distinct roles that are part of job descriptions and play a part in assuring the datacenter's processes. These roles have been chosen and separated because there are predefined patterns of communication hardwired into the roles. Thus, by looking at and defining patterns of communication, we can consider where roles can be combined into a single job description.

The roles and their short descriptions include:

- **Datacenter director:** interfaces with CUAHSI governance, ensures datacenter missions.
- **Software developer:** creates and maintains software crucial to datacenter operation.
- **Project librarian:** maintains releases of software and executes test plans.
- **Service administrator:** maintains reliable data services.
- **Developer support engineer:** supports developers in making software available to the CUAHSI community.
- **User support engineer:** aids users in utilizing CUAHSI resources to accomplish research tasks.
- **Data catalog curator:** ensures that HydroCatalog contains up-to-date information.
- **Data source curator:** ensures that data sources provided by the datacenter contain up-to-date information.
- **Data product curator:** ensures that data products published by the datacenter are appropriately documented and useful to the community. This is a role that will substantively increase in importance later in the 5-year plan for the datacenter's operations.
- **Standards curator:** ensures that documentation on relevant standards remains timely and relevant.
- **Wrapper curator:** ensures that wrapper software continues to properly map between non-CUAHSI and CUAHSI metadata.
- **Outreach coordinator:** implements outreach plans to reach and train potential users.
- **Contract manager:** maintains the set of contracts between the datacenter and researchers that serve as revenue streams.

The envisioned primary patterns of communication between roles are depicted in the following diagram:



In the diagram, arrows represent primary communication paths. As in typical development projects, it is best for primary development roles to be protected by a triage boundary and exempt from front-facing tasks except in triage escalation situations. Three front-facing roles satisfy needs of users and external developers.

Software developers must also respond to requests from standards and data source curators, so those lines of communication are noted. Data source and catalog curators must exchange information about the state of services. The project librarian, by contrast, is the technical second-level triage, and interfaces with both the service administrator and the software developer in coordinating releases, service changes, and outreach activities. The purport of the diagram is that roles with lines between them would make sense to combine into one job description, while roles without lines do not reduce communication by being combined.

Staffing Plan

The initial staffing for the datacenter include the following job descriptions:

- One **Senior Software Engineer** who fulfills the roles of
  - Datacenter director
  - Senior software developer and principal architect.
- One **Junior Software Engineer** who fulfills the roles of
  - Junior software developer (Web Programming/User Interface).
  - Developer support engineer.
  - One Development/Operations (DevOps) Engineer who fulfills the roles of
  - Service administrator.
  - Project librarian.
  - Junior software developer (HydroServer).
- One **User Support Specialist** who fulfills the roles of
  - User support engineer.
  - Outreach coordinator.
  - Catalog, data source, and data product curator.
- Subcontracts to:
  - SDSC Geospatial Computing Group for standards and wrapper curation.
  - Idaho State (Dan Ames) for HydroDesktop user support (which the datacenter will not be modifying as part of this proposal).
  - Utah State University (Jeff Horsburgh) for HydroServer support.
  - Lawyer for advice on contracts. CUAHSI Business manager to administer contracts.

This plan is based upon the above role diagram. Curation and user support are intimately connected so that it makes sense to assign one person to all of these. For the short term, software development will proceed along three separate tracks (documented below), and thus three staff will share development responsibilities of overall architecture, data services, and user interface. The staff member responsible for data services will double as librarian and service administrator. The staff member responsible for user support will double as data curator.

In year four, the junior software developer will be replaced by another curator, whose title is "Data Analyst", and whose responsibilities include creating data products to serve community needs.

Preliminary Budget

The preliminary budget for the datacenter is as follows:

	Year 1		Year 2		Year 3		Year 4		Year 5		
Item	w/o fringe	total	w/o fringe	total	w/o fringe	total	w/o fringe	total	w/o fringe	total	Comments
Personnel											
Senior Software Engineer	$ 80,000.00	$108,000.00	$ 82,400.00	$111,240.00	$ 84,872.00	$114,577.20	$ 87,418.16	$118,014.52	$ 90,040.70	$121,554.95	Software design and development
Junior Software Engineer	$ 60,000.00	$ 81,000.00	$ 61,800.00	$ 83,430.00							Transitional role
DevOps Engineer	$ 60,000.00	$ 81,000.00	$ 61,800.00	$ 83,430.00	$ 63,654.00	$ 85,932.90	$ 65,563.62	$ 88,510.89	$ 67,530.53	$ 91,166.21	Librarian and system administrator
Data Analyst					$ 60,000.00	$ 81,000.00	$ 61,800.00	$ 83,430.00	$ 63,654.00	$ 85,932.90	Evolving role
User Support Engineer/Curator	$ 60,000.00	$ 81,000.00	$ 61,800.00	$ 83,430.00	$ 63,654.00	$ 85,932.90	$ 65,563.62	$ 88,510.89	$ 67,530.53	$ 91,166.21	User support and curation
Subcontracts											
Director -- 1/2 time, 1 year		$103,732.00									Transitional role
SDSC (Standards and Wrappers)		$ 50,000.00		$ 50,000.00		$ 50,000.00		$ 50,000.00		$ 50,000.00	Wrappers and standards
Dan Ames (User Support/HydroDesktop docs)		$ 25,000.00		$ 25,000.00		$ 25,000.00		$ 25,000.00		$ 25,000.00	User support
Jeff Horsburgh (HydroServer support/docs)		$ 25,000.00		$ 25,750.00		$ 25,000.00		$ 25,000.00		$ 25,000.00	HydroServer support
Infrastructure											
Computers		$ 6,000.00						$ 6,000.00			Equipment replacement every 3 years
Networking		$ 1,080.00		$ 1,134.00		$ 1,190.70		$ 1,250.24		$ 1,312.75	Source: Verizon premium business servic
Cloud infrastructure services (upper bound)		$ 4,044.00		$ 4,246.20		$ 6,840.00		$ 7,318.80		$ 7,831.12	Source: Amazon EC2, non-elastic version
Cloud search services (lower bound)						$ 55,566.00		$ 58,344.30		$ 61,261.52	Source: Amazon Map/Reduce
Total Direct Cost		**$565,856.00**		**$467,660.20**		**$531,039.70**		**$551,379.63**		**$560,225.66**	
Indirect Cost		$374,483.50		$309,497.52		$351,442.07		$364,903.04		$370,757.34	
GRAND TOTAL		$940,339.50		$777,157.72		$882,481.77		$916,282.66		$930,983.00	

Notes:

- These figures do not include revenue predictions from subscriptions.
- Salaries include 35% fringe benefits and a cost-of-living raise of 3% per year.
- Equipment includes replacement of local datacenter equipment every 3 years.
- Cloud costs are upper bounds based upon commercial-level service.
- Network cost is an upper bound for commercial-grade networking.
- Network and cloud costs include projected 5% increase.
- Datacenter director on ½ time for one year as transitional measure.
- Subcontract to SDSC is ½ FTE.
- Subcontracts to Dan Ames and Jeff Horsburgh are ¼ FTE each.

The staffing plan above is a bare minimum that allocates too few programming staff for a timely initial transition. Utilizing front-facing staff to accomplish programming tasks (e.g., the User and Developer Support Engineers) will lead to substantial loss of programming productivity. In general, the roles are overloaded and the organizational chart lacks sufficient organizational maturity according to industry standards for services. We hope that these issues will be addressed by implementing revenue sources, enabling the datacenter to reduce its dependence upon external funding and hire more staff in years 3-5.

Computing Infrastructure

At the strategic/design level, hosting of CUAHSI HIS services is an ideal application of cloud data storage, positioned well to take advantage of all of that paradigm's strengths and avoid its known weaknesses. Attributes of CUAHSI HIS that make it already particularly suitable for cloud deployment include:

- Minimal need for user privacy. Almost all data is public. The only exception is that of private data sets tagged for delayed release, e.g., after papers are published.
- Minimal risk of impact from accidental disclosure of private data. Thus, weaknesses in cloud privacy do not significantly affect the project.
- Stateless information delivery, horizontally scalable via standard mechanisms for service elasticity.
- Relatively simple download and search services that can be re-implemented straightforwardly on a variety of cloud platforms.

In its current state, CUAHSI HIS services are suitable for deployment on Infrastructure-as-a-Service (IaaS) clouds that support Microsoft .NET. The services are already horizontally scalable in that environment, in the sense that there is no assumption of continuity or state in a sequence of service calls.

With some effort at adaptation, CUAHSI HIS services could be made suitable for deployment on Storage-as-a-Service (SaaS) clouds, or even Platform-as-a-Service (PaaS) clouds. This adaptation would allow CUAHSI HIS services to:

- Escape the performance limits of Structured Query Language (SQL) and database management systems (DBMS) including the MSSQL DBMS platform on which they are currently implemented.
- Employ NoSQL data storage methods and parallel NoSQL searches in selected cloud environments.
- Decrease response time for complex queries by a couple of orders of magnitude.
- Employ searches based upon at-scale search of data content – rather than just metadata – which are not feasible with the current (non-cloud) computing infrastructure.

Tactical recommendations

Contract a cloud storage service to host CUAHSI HIS and its descendants:

1. Define service expectations.
2. Explore strategic partnerships with Microsoft (Azure), Google (AppEngine), EMC (BigData), and Amazon (EC^2).
3. Develop a transition plan from the current implementation of CUAHSI HIS services to a version that exploits full cloud advantages.
4. Utilize the datacenter role of software developer to translate CUAHSI HIS services into a form suitable for the cloud.

TRANSITION PLAN

The above tactical plan represents steady-state and sustainable behavior that is the targeted result of a transition plan. The transition plan requires several elements, including:

- Software development:
    - Re-engineering and documentation of selected elements of CUAHSI HIS, including:
        - HydroServer data publication services.
        - HydroCatalog data discovery services.
        - All current service wrappers.
    - Development of workflow-enabling software for use inside the datacenter, including:
        - Curation interfaces for HydroCatalog, service wrappers, and data sources.
        - Backup and failover services for WaterOneFlow-compliant data sources.
        - Transition software for transitioning from remote to local data publication.
    - Development of front-facing software for use outside the datacenter, including
        - A highly accessible and generally usable web-based data discovery interface.
        - Data entry and upload software in support of datacenter data source hosting and publication.
        - Data usage reporting software for data providers.
- Outreach activities to promote use of datacenter data management include:
    - Deprecation of HydroServer solution and outreach to its current users.
    - Outreach to small-scale research projects with no permanent data publication plan.
- Policies to be crafted and approved include:
    - Privacy statement.
    - Statement of limitation of liability.
    - Data management policy that describes which management tasks the datacenter will and will not contract to provide.
- Reusable legal contracts to be written and developed include:
    - Subscription agreement for data hosting.
    - Subscription agreement for wrapper hosting and maintenance.
    - Subscription agreement for user support.
    - Subscription agreement for developer support.

Transition tasks and priorities:

The following tasks must be accomplished to set up the center, in order of priority:

1. Host HydroCatalog inside the datacenter's cloud (2 months programming).
    a. Re-engineer HydroCatalog code for robustness and add comprehensive documentation.
    b. Define and implement comprehensive testing for HydroCatalog services.
    c. Define and implement release management based upon that testing.
2. Define, implement, and support a curation workflow for HydroCatalog and data sources (2 months programming).
    a. Add curation components to HydroCatalog, including metadata that indicates that a source is permanently or temporarily unavailable, and metadata for redirection to an alternative source.
    b. Create a straightforward user interface to curation.
    c. Create a curation workflow and schedule for regular catalog maintenance.

    d. Define a curation workflow for wrapper data source maintenance.
3. Define, implement, and support cloud backup and failover for existing HydroServer and other WaterOneFlow instances (2 months programming).
    a. Re-engineer HydroServer code for robustness and add comprehensive documentation.
    b. Define mechanisms for cloud backup of WaterFlowOne service instances.
    c. Periodically check function of all service instances and annotate HydroCatalog.
    d. Define restore services and failover mechanisms for HydroCatalog in case of contingencies.
4. Create, document, and support a simplified desktop client targeted at beginning users (2 months programming).
    a. Utilize web technology and aim for broad platform support (Windows, Mac, Linux, Android, IPad)
    b. Utilize simplified GIS (e.g., rectangles) and faceted search.
5. Create virtual HydroServer instances for researchers who do not wish to run their own servers (2 months programming).
    a. Virtualize HydroServer into the datacenter cloud.
    b. Define, implement, and support an API and user interface that allows complete remote management by the researchers.
    c. Implement transitional software that creates a virtual HydroServer instance from a real one.
6. Implement social tracking of search metadata (2 months programming).
    a. Implement user-based tagging and tag sharing for HydroCatalog data.
    b. Add tagging and tag-based searching to simplified desktop client.

Note that many of these are programming-intensive, and that many can be done in parallel.

Of the tasks above, (1), (2), (6) represent HydroCatalog changes, (3) and (5) represent HydroServer changes, and (4) is a user client. Thus three programmers can be employed to accomplish these tasks in a rather straightforward way without conflict or undue dependence between the subprojects.

Based upon the staffing diagram above, and assuming that the DevOps and Junior Software Engineers are only ½ time programmers, rough estimates of times for development of these components are:

Task	Staff	Completion time
Re-engineering HydroCatalog	Senior Software Engineer	2 months after start
Curation interface	Senior Software Engineer	4 months after start
Cloud backup and failover	Senior Software Engineer	6 months after start
Simplified desktop client	Junior Software Engineer	6 months after start
Virtual HydroServer instances	DevOps Engineer	6 months after start
User-based source tagging	Junior Software Engineer	8 months after start

These estimates may be too ambitious, depending upon startup times for each of these staff positions, and the ability to retain people with the unique skills necessary to fulfill each role.

Milestones

Draft project milestones include:

- Year 1: initial transition from CUAHSI HIS to datacenter operations.
    - Complete first re-engineering of HydroServer and HydroCatalog.
    - HydroCatalog moves to IaaS cloud in the first of two transitions.
    - Complete first revision of an accessible data discovery client.
    - Complete catalog curation software and curation workflow; catalog curation workflow process becomes functional.
    - Complete data hosting software: data hosting plans and wrapper management plans offered for first time.
- Year 2: introduction of first set of advanced services.
    - Backup, restore, and failover services become functional.
    - Data hosting and failover subscriptions are offered for first time.
    - HydroServer deprecation and hosting alternatives announced.
    - Direct-to-cloud publication options for vendor hardware announced.
- Year 3: introduction of second set of advanced services:
    - HydroCatalog and Virtual HydroServer transition to hosting in PaaS cloud; advanced (data-intensive) search services announced.
    - Turnkey transition from remote to datacenter-based data hosting becomes functional.
    - Cloud-based hosting services are announced.
    - Internal datacenter processes are automated, including conversion from private to published data, consequences of subscription expiration, etc.
- Year 4: transition to sustainable activity levels:
    - Downsize programmer staff to sustainable levels.
    - Commit to operational policies for sustainable operation.
    - Automate processes to replace staff.
    - Upsize curation staff and institute development, creation, publishing, and curation of community-driven data products.
- Year 5: data product contracts available.
    - Datacenter enters a steady-state mode, with primary function to curate data.

Of course, this does not include any technological advances that might become part of the datacenter's activities during this time.

Possible Future Directions

In this document, we have considered only services that are currently already deployed or on the immediate horizon. Once these are in production, the datacenter may branch out in any of the following future directions:

- Providing comprehensive data management and data sharing functions for funded researchers.
- Forming strategic partnerships with companies interested in supporting data management.
- Federating CUAHSI HIS with other upcoming GIS data catalogs.
- Supporting source formats other than time series.
- Creating standards-based data products for community use.

- Providing "data fusion services" that combine and cross-validate time series data against other sources.

These are governance decisions that it is best to make once the datacenter is operating.

OTHER CONSIDERATIONS

Location of the Datacenter

The strategic decision to employ cloud technologies to host the datacenter's data impacts where the datacenter will be located. It is not necessary to locate the datacenter near a technology nexus; the computing infrastructure can mostly be remote, and the entire infrastructure can be managed via remote access. It is desirable, however, to locate the physical datacenter near a nexus of expertise in hydrology; it is quite crucial to involve domain experts in hydrology in the curation of standards, data sources, and data catalogs. These experts need not be faculty or researchers in hydrology; many tasks within these roles are well within the capabilities of a literate graduate student. Thus, a potential win-win situation is for the datacenter to hire graduate students enrolled in existing hydrology programs to assure the domain-specific workflows of the datacenter.

There is still a need for a physical datacenter; several of the communication paths in the above map of roles are too intensive to justify full remote management. In particular, it makes sense for the programmers and user- and developer- support personnel to occupy the same physical site.

Selection of Datacenter Director

The main other tactical consideration that has been raised is whether the director of the eventual datacenter must be hired immediately, or whether that hiring can be deferred until after the initial transition from CUAHSI HIS under an acting director. The advantages of hiring a director immediately include:

- Guaranteed continuity during the multi-year transition plan.
- Datacenter location can be tuned to the needs of the director, thus making recruitment easier.

However, the datacenter as designed is a community-governed effort. The director does not govern strategic or tactical directions; the director serves instead as a technical resource as to the feasibility and cost of a direction. In the current plan, the senior software engineer holds this role. The strategies and tactics for the datacenter are already determined in some detail. Thus the director's role is not central to the success of the datacenter; the actions of the CUAHSI Board are more crucial. The advantages of undertaking a search for a director during the first year include:

- Lower time pressure and a more comprehensive search, leading to higher-quality candidates.
- It is easier to secure a director after funding is assured.

However, getting the datacenter up and running will take much longer without an acting director in place. The recommended strategy is to seek a director who also functions as a senior software engineer and architect, and oversees the whole process of re-engineering.

Organizational Maturity

Another issue for a service-centered datacenter is whether the datacenter has a critical mass of employees with which to provide sustainable levels of service. In the initial phases, organizational maturity is not as much of a concern, but when the datacenter gets to year 4 or so, it becomes crucial. Organizational maturity issues include:

- How well are personnel roles documented, so that others can learn to take existing roles?
- How well can personnel change roles, in the case of personnel turnover?
- How much protection is there against catastrophic turnover, in which there is a multi-month period spent training new personnel?

While not an immediate concern, design of the datacenter's staffing to minimize the effects of turnover will be an issue in coming years. In such a datacenter, institutional memory is an issue and turnover is a constant issue. Thus the datacenter's job descriptions must be designed to minimize the effects of turnover. The datacenter does not have many paths for internal advancement, so turnover cannot be discouraged.

DATA SOURCES AND ACKNOWLEDGEMENTS

This report was compiled from information from many sources, including

- Regular interaction with the HIS teams over several months.
- An in-person visit to the SDSC Geospatial Computing Group.
- Several detailed tactical reports on potential operational policies and operating models for the datacenter, by Jeffrey Horsburgh, Utah State University.

REFERENCES

1. UniData, Unidata 2020: Geoscience at the Speed of Thought, December 2011.
2. Witold F. Krajewski et al, "Towards better utilization of NEXRAD data in hydrology: an overview of Hydro-NEXRAD," *Journal of Hydroinformatics* 13(2), 2011.
3. Anton Kruger, Witold F. Krajewski, Piotr Domaszczynski and James Smith, "Hydro-NEXRAD: metadata computation and use," *Journal of Hydroinformatics* 13(2), 2011.
4. Bong-Chul Seo, Witold F. Krajewski, Anton Kruger, Piotr Domaszczynski, James A. Smith and Matthias Steiner, "Radar-rainfall estimation algorithms of Hydro-NEXRAD,", *Journal of Hydroinformatics* 13(2), 2011.
5. Roger S. Pressman, *Software Engineering: A Practicioner's Approach*, Seventh Edition, McGraw-Hill, 2009.
6. Frederick Brooks, *The Mythical Man-Month: Essays on Software Engineering*, Anniversary Edition, Addison-Wesley, 1995.
7. Tom Limoncelli and Christine Hogan, *The Practice of System and Network Administration,* Addison-Wesley, 2002.
8. Daniel Menasce and Virgilio Almeida, Capacity Planning for Web Services: Metrics, Models, and Methods, Prentice-Hall, 2002.

APPENDIX A: FROM PROTOTYPE TO PRODUCT

One goal of the datacenter is to transform selected CUAHSI software development efforts from "prototype" to "product". In software engineering, a "prototype" is an experimental version of software that concentrates upon "what it should do"; a "product" or "production service" also contains mechanisms for software quality assurance, including comprehensive documentation, test plans, and release strategies, among others[6].

While these terms traditionally refer to software, they are equally applicable to services currently provided by the CUAHSI community on a best-effort basis. The "prototype" in that case is a service provided by CUAHSI volunteers, while the "product" is a service provided by the datacenter. Transforming CUAHSI HIS into a "product" requires engineering these services so that they are provided reliably and not simply on a best-effort basis.

To accomplish this, the datacenter adds an engineering layer between prototype and product, similar to that in the development process for other open-source projects. The datacenter empowers hydrologists to extend the software, and supports their efforts by ensuring the quality of the result. Ideally, the hydrologists define the "requirements" via prototyping, while the datacenter provides the resulting "product". This is a rather unique model of development that we believe will produce unique results and advance the state of the art.

This can be expected to be a costly process, however: the software engineering literature estimates that there is roughly a factor of 10 difference between the development expense for a prototype and a product[5,6]. Doing this in the context of a datacenter absorbs this cost within the expertise of the datacenter, rather than attempting to place this burden upon the scientist-developers.

The External Software Developer's View of the Datacenter

From a developer's point of view, the steps involved in interacting with the datacenter are as follows:

- The developer conceives of and prototypes a feature.
- When the developer is satisfied that this feature should become permanent, it is submitted to the datacenter as a potential feature to be supported by the datacenter.
- A governance decision is made as to which features the datacenter will support in production, and which will remain in prototype form. This will be done by an oversight subcommittee of the CUAHSI board.
- Once a feature is accepted for production, the datacenter collaborates with the developer in making the changes that support production use, and then takes over maintenance of the component, in perpetuity, even if the developer is no longer interested in supporting the component.

In order to fully understand the role of the datacenter in software development, it is necessary to describe in detail their interactions with other software developers during the whole software lifecycle. The typical lifecycle of a software project consists of several distinct phases, often called "the waterfall model" [5]:

1. **Requirements analysis:** determining what the software should do.
2. **Design:** determining how to accomplish requirements.
3. **Implementation:** writing code according to the design.
4. **Testing:** determining whether code accomplishes requirements.
5. **Maintenance:** reacting to requests for changes, problems, etc.

So far, the scientist-developers have been responsible for all of these phases.

The datacenter will support a modified lifecycle based upon the assumption that hydrologists will contribute prototypes to be turned into products. The scientist-developer will engage creating a prototype (or proof-of-concept) via:

1. **Requirements analysis:** determining what the software should do.
2. **Design:** determining how to accomplish requirements.
3. **Prototyping:** implementing an initial version of the design.
4. **Initial testing:** proving the concept that requirements are met.

This is what the CUAHSI scientist-developers are already doing. In these tasks, the developers can obtain (optional) technical assistance from the datacenter.

The key to understanding the science/engineering boundary in this process is to make a clear distinction between the technical software engineering processes of validation and verification[6].

- **Validation:** "Are we making the right product?" Do the product specifications satisfy user needs?
- **Verification:** "Are we making the product right?" Does it conform to specifications?

In these terms, the scientist-developer completes a validation of the prototype, and then turns the resulting prototype over to the datacenter's engineers for product engineering and ongoing verification.

The next phase in the prototype lifecycle might be called **admission.** In this phase, the prototype is transformed into a product suitable for ongoing maintenance.

1. The developer applies to include the prototype as part of the CUAHSI software product.
2. A governance decision is made by the CUAHSI Board as to which prototypes to include.

If the application is successful, roles change between developer and datacenter engineers. The engineers take the lead in the following engineering processes, in close collaboration between developer-scientists and datacenter engineers:

1. **Documentation:** documenting aspects of the prototype that are needed by the engineering team or end-users.
2. **Re-engineering:** changes in the requirements, design, and implementation of the prototype needed for effective production use.
3. **Test planning:** establishing the workflow for testing the resulting product or product feature before each release.

Once these processes are completed, the prototype becomes a product, and the final **maintenance** phase begins. In this phase, datacenter engineers maintain the product in isolation from the scientist-developer, including:

1. **Software maintenance:** reacting to bug reports and user requests, as well as incorporating changes requested by the original developer.
2. **Change management:** tracking the changes and reasons for changes in the product over time.
3. **Release engineering:** ensuring that each software release is appropriately tested and verified.

Once a component is being maintained by the datacenter, the original developer can still request changes; these changes are scrutinized for problems by the engineering team before the changes are released. But there is no necessity for further scientist-developer involvement at all. Components written by student software developers continue to be maintained long after the students have graduated.

Some parts of the CUAHSI product are so tied to standards efforts that the communication between scientists and engineers need not concern software at all. HydroServer and HydroCatalog are best viewed not as software products, but instead as reference implementations of standards. The inner workings of these products are esoteric and unrelated to water science. Providing prototypes of software to describe new standards is an inefficient way to specify the standards. It is best for the scientists to propose standards that the engineers then implement; development involvement by the scientists is neither desirable nor empowering for the engineers in embodying the standards.

So far, the engineering team at the Geospatial Computing Group of the San Diego Supercomputing Center has functionally served as the "software engineers" for the CUAHSI product. They have taken control of the HydroCatalog service and made it an embodiment of standards. This is proof that such science/engineering collaboration can be fruitful, but engineering the whole CUAHSI product is beyond the SDSC group's primary missions.