NGDA – National Geospatial Digital Archive

An interim project review report

Reviewer

Dr. Philip J. Sallis
Professor of Computer Science
Auckland University of Technology
Aotearoa-New Zealand

and

Davidson Trust Research Fellow, 2005-6
Map and Imagery Laboratory
UC Santa Barbara

Acknowledgement

The author acknowledges with gratitude the funding made available from the award of a Research Fellowship Grant provided by the Davidson Library Trust, which enabled the work for this report to be carried out.

Introduction

This report is an overview of the project to create a **National Geospatial Digital Archive**. It is an interim report in the second year of the project, as of February 2006. The review has been based on internally published documentation concerning the project and its progress.

Funding for this project (approximately $15 million) is provided by The Library of Congress. UCSB and Stanford University are jointly receiving approximately $2.8 million for an initial three year period to carry out collaborative research and development work associated with creating a technical architecture concept for the repository (the '*Archive*') and its data administration.

This review acknowledges that a three year research and development project will produce a working prototype of the NGDA concept but that a full production system for universal use will require a longer period for development and data migration. This needs to be clear to the project's sponsors, researchers, developers and users alike. To evolve a fully operational NGDA will also require a dedicated production team appropriately titled and located. This point is made here to indicate that the initial research and development funding provided will

require supplementation for both the next stage of system construction (from prototype to fully operational system) and on into the future for the production environment to function. The question of whether this is likely to be fully user funded, partially user funded or totally subsidized by the Library of Congress or other agency is not evident form the documentation available to this reviewer.

There has never been a common geospatial format registry in production, other than for text and common image items. The need for such a registry is both to create a single repository for geospatial data (graphic, textual and numeric) and to establish a 'timeless' meta-data set that will 'future-proof' the formal definitions and descriptions of the geospatial items.

The project value proposition requires formal specifications to be developed for the content of the *Archive*. It also requires the production of formal software specifications for programs to be written that will format and migrate the geospatial data that has been acquired from various sources. Formal specifications for the interface software (system and user) are also required.

The common format registry will be used for this purpose. The items will thus be ingested into the *Archive* and otherwise processed. A temporally enduring form and content architecture is required, in order to anticipate future changes to computational methods and data specifications.

The reviewer observes that the prototype being developed here, in common with all geospatial data processing applications, is a non-trivial system with complex data structures and large sets of data. This characteristic of the project should not be understated or its potential failure points under-estimated.

Ensuring robustness of the system architecture, with high quality precision specifications and a thoroughly tested data model, which is pivotal to the uniqueness of the design proposed for this prototype, is essential as a foundation for quality assurance. In terms of risk assessment, this early formal specification task is generally regarded in the field of Software Engineering as being of the highest priority.

The following text in **Figure 1**, is an extract from what is published on the original UCSB-Stanford Team project website. It outlines the perceived need to establish a **National Archive of Geospatial Information**. It also describes the NGDA project and its objectives. The author of this report has used this set of value proposition statements as the essential benchmarking point for the activities and outcomes of work carried out so far, as it is documented. All of what has been detailed for the work-in-progress has then been benchmarked against contemporary industry standards and international best practice and principles in the literature relating to the field of Software Engineering.

<Extract begins>

NGDA - National Geospatial Digital Archive

The University Libraries of UCSB and Stanford are leading the formation of the National Geospatial Digital Archive (NGDA), a collecting network for the archiving of geospatial images and data. Geospatial information has played an important role in the history of the United States. From the first colonial maps to the satellite imagery of the 21st century, cartographic information has helped define and frame our view of the United States. This will be done under a grant from the *Library of Congress*. Concerned that millions of nationally important digital information resources are in danger of being lost or corrupted, the *Library of Congress* has partnered with eight institutions to begin a three year $15 million effort to build a nationwide digital collection and preservation system. The Partners will seek additional collaborators from university, government, professional society and corporate sectors through the life of the grant and beyond. We would encourage those who have collections of geospatial data to contact us regarding participation.

The objectives of our project are to:

- Create a new national federated network committed to archiving geospatial imagery and data.
- Investigate the proper and optimal roles of such a federated archive, with consideration of distant (dark) backup and migration, directly serving content to users, vs. referring requestors back to the originators of the data for copies or assistance, active or passive quality/integrity monitoring, application of metadata, federated searching, dissemination of metadata, etc.
- Collect and archive major segments of at-risk digital geospatial data and images.
- Develop best practices for the presentation of archived digital geospatial data.
- Develop partner communication mechanisms for the project and then ongoing.
- Develop a series of policy agreements governing retention, rights management, obligations of partners, interoperability of systems, exchange of digital objects, etc.

<Extract ends>

Figure 1 - Project Description and Objectives

The Documentation

All of the project documentation is either stored or referenced at the *Confluence* project website. This is a secure website for access by those associated with the project and acts as a focus for activity recording.

The 'official' project information website at www.ngda.org appears somewhat 'clunky', considering contemporary website design conventions and the building tools available. In contrast, the original website containing the overview and objectives information in **Figure 1** above, which is to be found at www.alexandria.ucsb.edu/~masi/ngda, is of contemporary design, is informative and is appealing to read.

Project documentation resides at or is linked to from the *Confluence* website at www.alexandria.ucsb.edu/confluence/display/NGDA/Technical+Architecture.

It is clear from the website content that there are weekly project meetings held, which are documented and from those meetings, items that are appropriately actioned. The **First Year Roadmap** [also at the *Confluence* website] provides a description of concepts and ideas that correlate with the 'objectives' outlined above. This paper is both conceptual in terms of overall philosophy and system design and also technical in its description of data management services that are currently available in web-based products and special-purpose programs.

Other papers and project notes are located here that relate to for example, a technology scan that was carried out in order to chose the best software development tools and methods available for use on this project. Products such as *DSpace* (www.**dspace**.org) for rapid prototyping and *Fedora* (www.**fedora**.info) for digital objects storage processing, together with other existing development tools were considered. While these open source products are appealing in many application situations, and they may be utilized later during the project evolution, apparently none of them completely satisfied the requirements for this project. It was considered more appropriate to use a hybrid of tools including some that have already been developed for use with the Alexandria Digital Library (**ADL**).

So the results from this evaluation led to the unique approach and combination of methodologies and technologies described in the technical notes at the *Confluence* website, including descriptions of the Archive Format, XML Manifests and Templates to be used, together with the Data Model and System Architecture.

When comparing the extant documentation for the project (at the *Confluence* website) to the set of project objectives outlined in **Figure 1** above, it is clear that each of these is being addressed, although at differing stages of development. A report that outlines project goals to be achieved by August 31, 2006 makes this

correlation apparent, and progress to date can be compared with the items in **Figure 1** as in the points below:

- A national federated network is being evangelized and participants are being located.
- A data model and system architecture have been proposed and established, with data already populating the *Archive* for retrieval via both the **ADL** platform and a special-purpose interface currently under development.
- Two collections of 'at-risk' data have been located and are being ingested to the *Archive*.
- The environmental scan, which produced the architecture and data model solution now being adopted, is contributing to best practice for scalable data storage and retrieval systems.
- The project teams are meeting, corresponding and mutually benefiting from the collaborative nature of the project.
- Policies and processes are either developed or under consideration as can be seen by reviewing the documentation located at the website.

The Paper, **NGDA Second-year Goals and Tasks** at the *Confluence* website describes five facets of the project that have goals, rationales for the goals (asking the question "Why do it?"), and numerated objectives for this second year. Addressed here are matters such as awareness raising in libraries and among the geospatial research community, formal specification for a registry of format specifications, establishment of the *Archive* itself, repository content and issues of access. These all correlate with the published statement of project objectives.

It appears then, from a cursory analysis, that the project is well directed and managed. It is on schedule with work-in-progress in relation to the interface software and meta-data translation processes, together with an established ingestion method that has been conceived and is now building up the *Archive*.

System Architecture

The Data Model is described in a Project Note [see NGDA Archive Data Model] at the *Confluence* website previously cited at www.alexandria.ucsb.edu/confluence/display/NGDA/Technical+Architecture.

This project note refers to the internal system representation descriptions, the types of data directories to be created, the inter-object relationship definitions and the schema specifications. When considering this data model in terms of the

overall System Architecture, it is useful to view the Operational Specification diagram replicated below in **Figure 2**.

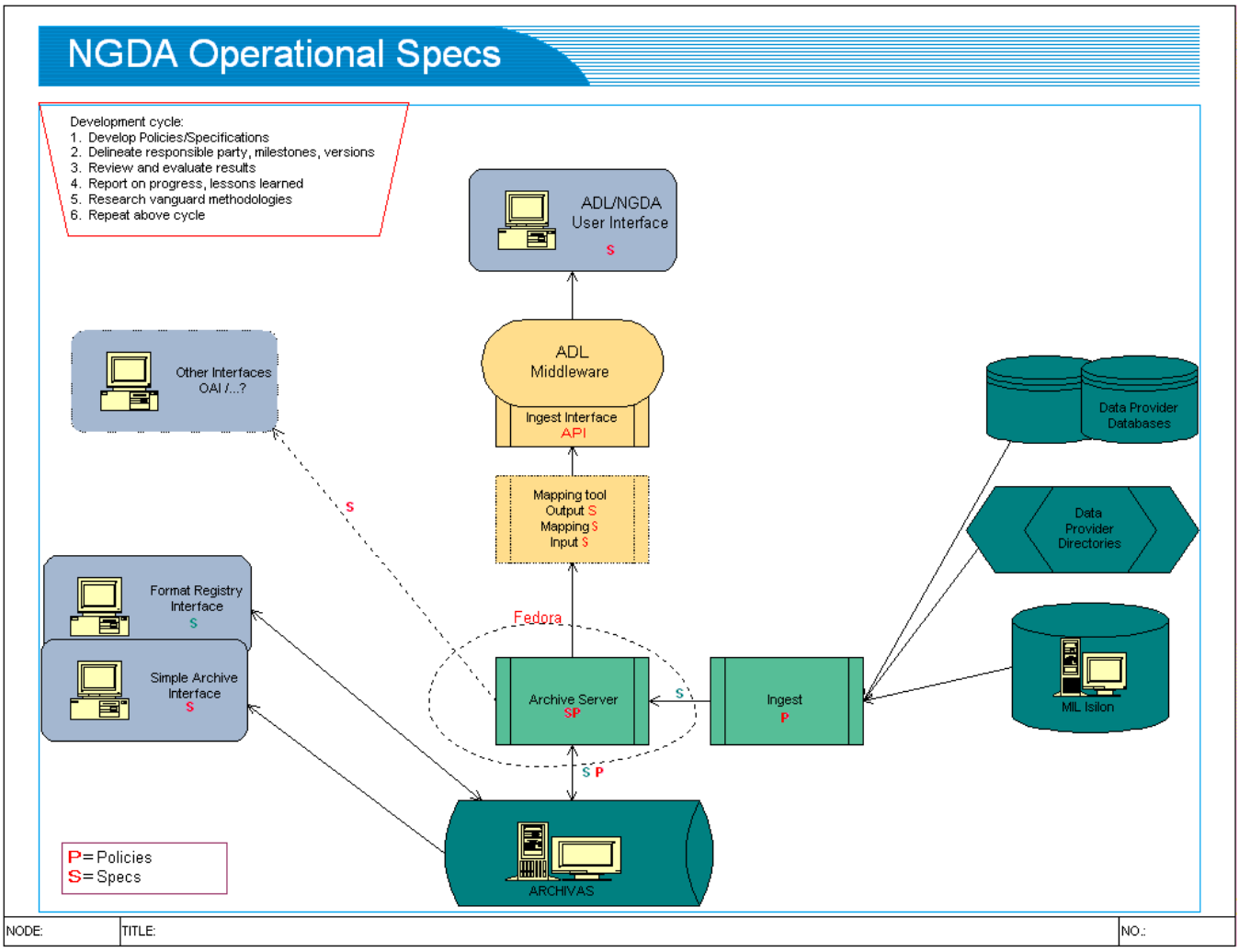

Figure 2 - **Operational specification for the NGDA project**

The specification diagram clearly shows the need to accept data in any format and transform its meta-data into a form that is both consistent with **ADL** and the Format Registry Interface currently being developed.

Providing access through the **ADL** browser immediately provides a user market that will assist in the promulgation of the *NGDA Archive*, give access to a wider collection of materials, and enable/encourage use of it by a wider population of researchers and other interested parties. This in itself should assist in the library

education goal cited in the **Goals and Objectives for August 31, 2006** paper referred to earlier in this report.

A decision has been made to store all geospatial data items in the *Archive*, rather than merely referring users to other locations. Apparently this is both to ensure a complete and secure repository of items and thus achieve the primary focus of the project from the sponsor's perspective, and to have all items conform to a new single data record specification format, which ensures collection integrity and ease of system maintenance into the future. This approach is both conceptually sensible and strategically robust from a system architecture perspective.

The usability and usefulness of being able to retrieve any data directly from the *Archive*, either through **ADL** or the *Simple Archive Interface* (both depicted in the diagram at **Figure 2**) cannot be over-stated as a forward-looking and extremely efficient method for user acquisition of its content. In addition to the 'one-stop-shop' approach, which adds to retrieval effectiveness, this approach provides a very appealing potential user interface, cutting out mouse clicks, navigation, format and style changes, etc. It also ensures an enduring accessibility because as format and meta-data definitions change over time, the collection will be neutral.

The overall system architecture is to be commended on its thoughtful conceptualization. Its implementation appears to be well managed and on-track in terms of the timeline expectations recorded in the planning documents. Multiple locations of personnel for collaborative work can add a level of complexity to project management, document control and system administration. These elements appear to be well directed and effective for this project.

System Development

The choice of software technologies for development has been referred to earlier in this report. Both the templates used and the server protocols described in the project documentation are contemporary and appropriate for this application.

A decision has been made to use a hybrid of existing software development tools (XML Manifest and others) rather than simply adopt an integrated package approach such as using *DSpace* or *Fedora*. Given the special purpose software that has been developed for **ADL**, including some specific system and application tools, this decision is a wise one. Using ubiquitous software tools together with the **ADL** programs will provide a generic, yet context-relevant suite of programs. This has the benefit of having temporal robustness as well as being maintainable by the widest group of programmers.

As previously mentioned in this report, an interface with **ADL** both increases the interoperability of systems and extends the size of the overall digital collection. It also provides a systems and data integrity platform, which removes idiosyncratic formats and thus enhances both interoperability and accession potential by the widest possible community of users.

Finally, the decision to use ***Archivas*** (www.archivas.com) as the *Archive* back-end is a sensible one too. It is a proven product for the so-called 'dark' archiving required to ensure both the security of current data and its longevity in storage. The product appears to have excellent version control and data tracking features, important for the issues of security and integrity in the geospatial repository for this project. There is no obvious reason, from either a functionality or cost perspective, why this software should be developed locally.

Conclusions

A review of the documentation relating to the project at this interim stage of its evolution indicates that it is on schedule, is effectively directed and efficiently managed.

The architecture and underlying data model appears sound according to contemporary technologies and the choice of software development tools accords with industry standards for open source products. Together these provide maximum potential for data security and integrity, together with system interoperability and operational optimization.

The decision to create a common data record format and translate all sourced geospatial record descriptors into it is a very sensible future-proofing concept, and the ingestion of all sourced data into a single archive is similarly sensible for both future-proofing and usability.

The system design is in itself user-oriented and seeks to be optimized for access through the interfaces being built.

Having two user interfaces provides alternative access to the *Archive* and thus maximizes its use potential. This is particularly true of the access through **ADL**. User access to the *Archive* via **ADL** adds both the advantages of using existing data processing and information retrieval tools, while also extending the user's potential for seamless access to the entire **ADL** collection in association with the geospatial data in the *Archive*.

Finally, although there is no specific reference to how well the collaboration between UCSB and Stanford University is working, there does appear to be communication between the two research and development teams, with some face-to-face meetings having been held and joint decisions taken.

Recommendations

The first recommendation, from this interim review is that the website (at least the Homepage) be brought into line with contemporary practice in terms of industry standard usability characteristics and measures. These are widely published. This would enhance the professional appearance of the project and its management, especially given the project's application context. In this case, a 'make over' would be appropriate. Initial comments on this matter are made early in this report.

The second recommendation, is that the document tracking (reports, meeting notes, schedules and timelines etc) be better ordered according to a hierarchical classification scheme that clearly indicates dates, events and topics in the item identifiers. Additionally, that the format of technical reports, meeting notes and other insertions conform to a common template for presentation. This is non-standard through the documentation, although some order of the items and references is noted and the documentation is by no means disorderly.

The third recommendation, is based on the observation that this is a non-trivial system concept, with complexities of both software interaction and data inter-relationships. The operational environment needs to be both scalable and performance critical. Because of this reality, it is recommended that a set of performance monitoring metrics soon be identified and that some impact analysis be undertaken, probably using simulation techniques, to assess the operational impact of variables such as user demand via the interfaces, process throughput, query loads and database access, schema robustness and collection responsiveness.

The fourth recommendation, is that given the nature of this project being to produce a working prototype, once the *Archive* has been sufficiently populated, and the *Simple Archive Interface* has been developed, it be tested by an independent user group. Some attempt at formally mapping user profiles against the collection usage would be appropriate for the ongoing analysis of collection utilization and its concomitant system performance management.

The fifth recommendation, if it hasn't already been addressed elsewhere, is to articulate clearly the phasing of the prototype being developed within this segment of the overall NGDA Project and its outflow into the establishment of a fully operational production system. Producing a graphical representation of this evolution supported by some narrative to explain the contingencies, potential points of failure, timeline, physical and human resource requirements and cost estimation forecast is suggested as necessary for forward planning beyond the current research and development work detailed in this review report.

Summary

In summary, the author congratulates the governance group and management team on a methodologically contemporary, elegant and scalable system design, with a seemingly robust and efficient architecture, together with a data record format that should provide longevity for the geospatial items in the repository, which in itself is effectively administered by the archiving software being utilized.

Philip J. Sallis
Santa Barbara
February 22, 2006

The Author

Dr Philip J. Sallis is currently the Deputy Vice Chancellor at the Auckland University of Technology in New Zealand. He concurrently holds a Chair in Computer Science.

Philip was appointed full professor in 1987 while working at the University of Otago in New Zealand and before that held senior academic positions in Australia and the UK. He has held visiting research professorships in Hong Kong, the UK and at UCSB where he has spent one sabbatical leave in addition to several short-term visits since 1991. He was awarded a Davidson Trust Research Fellowship in 2005.

While at UCSB, Philip first worked with the NCGIA group and later with MIL. He has been associated with the Alexandria Digital Library during much of its development phase. He designed and conducted some system performance simulations, including user and collection utilization profiling, based on session log data analysis. This work utilized some contemporary mathematical modeling methods including computational neural networks.
This work was published in four papers jointly authored with ADL personnel. He has also conducted usability testing of ADL in the field and in 2003 established the NZADL…a regional ADL server, which is currently being used by researchers and graduate students.

Philip has published widely in the areas of Software Engineering and Computational Linguistics. He continues to speak at conferences, have membership of Government and business working groups and remains active in research despite his administrative responsibilities.