

Evaluation challenges for a federation of heterogeneous information providers: The case of NASA's Earth Science Information Partnerships

Catherine Plaisant[%], Anita Komlodi^{#%}, Francis Lindsay

University of Maryland

Institute for Advanced Computer Studies

[%]*Human-Computer Information Laboratory*

[#]*College of Information Studies*

plaisant@cs.umd.edu, komlodi@glue.umd.edu, flindsay@geog.umd.edu

Abstract

NASA's Earth Science Information Partnership Federation is an experiment funded to assess the ability of a group of widely heterogeneous earth science data or service providers to self-organize and provide improved and affordable access to an expanding earth science user community. As it is self-organizing, the Federation is mandated to set in place an evaluation methodology and collect metrics reflecting the outcomes and benefits of the Federation. This paper describes the challenges of organizing such a federated partnership self-evaluation and discusses the issues encountered during the metrics definition phase of the early data collection. Our experience indicates that a large number of metrics will be needed to fully represent the activities and strengths of all partners, but because of the heterogeneity of the ESIPs the qualitative data (comments accompanying the metric data and success stories) becomes the most useful information. Other lessons learned include the absolute need for online browsing tools to accompany data collection tools. Finally, our experience confirms the effect of evaluation as an agent of change, the best example being the high level of collaboration among the ESIPs which can be in part attributed to the initial identification of collaboration as one of the important evaluation factors of the Federation.

1. Introduction

Beside the obvious need to evaluate any experiment to measure its positive and negative impact the U.S. Government Performance and Results Act (GPRA) [1] is slowly changing the way federal projects are being conducted. Quantitative and qualitative metrics are being defined by projects but large and heterogeneous programs and experiments present a serious evaluation challenge.

The U. S. Government Performance and Results Act was created in response to complaints and findings of waste and inefficiency in federal programs stemming from 'insufficient articulation of program goals and inadequate information on program performance'. Among the goals of the GPRA is promoting a new focus on results, service quality, and customer satisfaction. It mandates the creation of strategic plans, including provisions for program evaluation. Annual performance plans and reports are another area covered by the Act. Each agency is required to create annual performance plans for each program. Goals need to be expressed in objective, quantifiable and measurable form.

GPRA calls for the establishment of 'performance indicators to be used in measuring or assessing the relevant outputs, service levels, and outcomes of each program activity'[1], and for the provision of a basis for actual program results with the established performance goals. It also requires a description of the means to be used in verifying and validating measured values. In cases where performance cannot be measured by objective and quantifiable terms, special arrangements can be made for alternative performance evaluation methods, such as descriptive statements of minimally effective and successful programs. The GPRA gives general definitions of output and outcome measures, performance goals and indicators.

In this paper we focus on one of NASA's Earth Science Enterprise experiments. The goal of NASA's Earth Science Enterprise (ESE) is to further develop our understanding of the total Earth System, including the effects of natural or human-induced changes to the global environment. The program draws upon the full range of NASA's technical and scientific capabilities to achieve this goal. The ESE distributes this information to both the public and the private sectors in order to promote productive use of the gathered data.

The Earth Science Information Partners Federation (or ESIP Federation) is an experiment – or working prototype – funded to assess if such a Federation could be the viable enterprise model to facilitate the public availability of Earth science data from geographically dispersed providers, and to foster the development of new services to the user community. The Federation itself is charged with collaboratively developing this new enterprise model. It is charged to define a governance structure that encourages cooperation as well as create a community of ESIPs that are dedicated to the continued success of the Federation.

The most striking characteristic of the Federation is the diversity of its partners:

- 10 “Veteran” Distributed Active Archive Centers (DAACs) who have been operational for years and deal with the archiving and distribution of Earth science data (called ESIP Type 1s).
- 12 new data-enhancing ESIPs, generally run by Earth scientists themselves at Universities. They create new products and services for the research community (called ESIP Type 2s). Examples of ESIP2s include the University of Maryland ESIP specializing in new data products and services for land cover studies, and the DODS ESIPs offering data interoperability software components.
- 12 commercial entities, from startup companies to established educational entities like museums, provide practical applications of Earth science data for a broader community (called ESIP Type 3s – they are only partially funded by NASA, and meant to become rapidly self sustaining). Examples of ESIP Type 3s goals vary from developing educational materials to developing new services for commercial shipping and fisheries.

This paper describes the challenges facing the ESIP Federation in organizing its own evaluation, and discusses the issues encountered during the metrics definition phase of the early data collection.

2. The Role of Evaluation

Marchionini [2] differentiates evaluation as a research process from product or system testing. Evaluation research examines the interaction of complex phenomena using qualitative and quantitative methods, studying the object of the evaluation from many different angles. The current evaluation aims at similar methods, as the application area and the organizational experiment are both new and complex phenomena.

The object of the current evaluation is a web-based information system. Methods of evaluation

suggested for more traditional information systems can be applied, although the technology is different and this should be taken into account. In both cases, information is the commodity that is collected and processed and provided to users. The overall goal and mission is the same for both, thus evaluation methods will have to be similar. Lancaster [3] suggests defining inputs, outputs, and outcomes in information system evaluation. The long-term objective of any such system is to achieve certain outcomes in the community it serves. Inputs are processed in order to generate outputs, in the case of information systems the goal is to turn financial input into information services output.

Griffiths and King [4] emphasize the importance of measuring not only performance, but also user satisfaction, purpose of use, consequences of use, etc. They suggest four kinds of measures: input costs, outputs, effectiveness, and domain. They define information system functions as user-related, operational and support functions. The four measures are collected in all these functional areas. In these terms, the evaluation reported here focuses on output, outcome and impact measures in user-centered functions.

Marchionini [5] provides one of the best examples of long term digital library evaluation. An evaluation of the Perseus Project, an evolving digital library of resources for the study of the ancient world was carried out. The evaluation was extensive and covered several distributed sites with different practices but they were all using the same multimedia system and the data could be collected fairly homogeneously. The author describes methodology, results and implications.

In the domain of geographical information systems Hill [6] reports on the evaluation of a single site and how it improved usability and user satisfaction. Other digital library evaluation discussions can be found in the D-Lib Magazine article of July/August 1998 [7], and in Bishop et al. [8].

3. The Earth Science Information Partnership Federation

One challenge for the ESIP Federation is the prototyping of alternative ways of developing, producing, and distributing Earth science data. The Internet, the World Wide Web, and other rapidly developing technologies enable the public to access vast quantities of information. Developing practical applications of advanced information technologies such as these is vital to the discipline of Earth System Science. Current research in the Earth sciences has the potential to yield a variety of new scientific insights and practical benefits - from monitoring

deforestation or understanding global environmental changes, to providing customized reports to farmers or fishermen, or offering services to local government and land owners. The ESIP Federation provides the means to facilitate data access, exchange, and enhancement.

The Federation was launched in 1998 with the competitive selection of the initial 24 ESIP Type 2s among candidates from the academic, government and private sectors.

ESIPs have different levels of funding and cost sharing, they are at different stages of developments – from the well established ESIP Type 1s, to others who are not to date fully operational. Some offer data only, other services but no data, and some both. Some ESIPs have a handful of employees, other several dozens. Some ESIPs easily reach thousands of users via a popular website (e.g. the NBC4 weather page) while others cater to a smaller number of high-end scientists who may use their service extensively. Finally another peculiarity of the Federation is that it will be evaluated by the level of cooperation among ESIPs. Individual ESIPs were selected following a competitive process and will have to re-compete to receive continuing support. Because of its self-organizing principle, the Federation receives limited guidance from NASA, which leaves room for experimentation but can also lead to confusion in the early stages of organization.

In simple terms the goal of the Federation can be described as the provision of improved and affordable access to Earth science data, and the development of new user communities attracted by new types of data and services.

4. Challenges of Evaluation

4.1 Leadership and Coordination in Evaluating Distributed Systems

The development of an evaluation plan, and its execution can be originated on different levels of a heterogeneous, distributed systems. It is important to clearly define responsibilities and coordinate actions among the different evaluation activities in order to minimize effort. An individual project can be collecting evaluation information for their own use, or collect data for reporting and management.

The development of evaluation metrics across all ESIPs will inevitably involve representation from these ESIPs. This is necessary to insure the inclusion of all the different types of activities and accomplishments of the different units. The management of this process was setup by NASA management decision and voluntary selection. The team at the University of Maryland took responsibility for defining and developing metrics,

and devising ways to collect them. In the development of the metrics, we built on several different evaluation efforts in the ESIP projects, and collected several existing sets of metrics. We defined a new set based on these and on our evaluation experience with our own ESIP, the Global Land Cover Facility (<http://glcf.umiaccs.umd.edu>). Another ESIP volunteered to use their existing web tools to develop prototypes of online data collection and viewing tools.

4.2 Including all aspects of what the partners do

The 34 ESIPs grouped under the aegis of the Federation differ considerably in their profiles. They vary in size, infrastructure, mission, and services. The heterogeneity makes it difficult to find a balance between (1) limiting the number of metrics, and (2) identifying metrics that truly represent the activities and strengths of all ESIPs.

Our strategy evolved over time. The first approach involved defining a small, common set of mandatory metrics that all projects would provide, later we added the notion of a larger group of optional metrics. We also considered including yet another level of metrics: custom metrics defined and collected by the individual ESIP in order to monitor and improve their own activities (e.g. what part of the website or dataset is most used? is the help system being used?).

This multilevel approach was rejected mainly due to the lack of agreement on a small set of common "mandatory" metrics. Instead, a single larger set of common metrics was developed, acknowledging that not all ESIPs will be able to collect all metrics and would just mark them as Non Applicable.

Examples: The initial list of metrics included measuring the amount of data delivered, the volume of data etc. because this was a traditional way of measuring ESIP Type 1s (DAACs) activities. Some of the ESIPs do not have any data per se, and instead provide valuable services (i.e. processing data provided by users, or providing interoperability capabilities and software components). In response we had to also measure the number and type of services.

One of the initial metrics was the total number of users, but in our first attempt to collect metrics data we realized that we were adding up science users downloading data, kids using education materials, and the general public reading web pages. It became clear that a better classification was needed to reflect the diversity and richness of the activities so we separated consumers (e.g. museum patrons, web page

readers) from data and service users (who download or order data).

Qualitative data. In order to accommodate the variety of the ESIPs, we collected both quantitative data (metrics) and qualitative data referred to as nuggets. Nuggets are short descriptions reporting on the special activities and accomplishments of the ESIPs. There are currently twenty categories of nuggets defined, based on discussions with the different projects.

Even after adding many metrics to our list of metrics there was still a strong feeling of not being able to completely capture the richness of activities and the nuggets have become the most important reporting method. For example one of the ESIPs did not have much data, very few users but generated results that could be better evaluated by the number of lives they saved! Clearly not a metric applicable to many ESIPs but could be reported via a nugget.

Examples of nugget types

New science:

- New type of data use
- Data quality achievement

Federation activities

- Notable results from working group
- Example of federation collaboration
- Collaboration with other institutions

Federation reactivity

- Rapid response to adverse event
- Rapid dissemination of data or service

Dissemination:

- Publications written by ESIP users
- Mention of ESIP in press

Education

- New K-12 education activities
- New higher education activities
- Student graduated

Miscellaneous

- Steps toward sustainability
- Impact on citizens, business
- Quotes

Examples of metrics:

- Number of data and service users
- Number of information consumers

- *Note: Consumers are all users who come to find information or learn from ESIP information but do not necessarily download data or use services*

Number of repeat users

Data volume

Total data volume in archives, including data not available to users

Number of Datasets

Number of Data Products delivered

Volume of products delivered

New datasets not available before the federation

Number of services available to users

Number of services that were not available before the federation

Number of services rendered

Delivery time of data or service

4.3 Dealing with parallel collection efforts

Because of heterogeneity of the ESIPs subsets of ESIPs are also part of additional metrics collection efforts. For examples the DAACs (ESIP Type1s) have been in existence for years and many metrics are already being collected. The more "commercial" ESIPs are part of a larger group of projects required to submit special metrics related to their economic impact or financial well-being. Those concurrent metrics collection efforts use different collection mechanisms (e.g. interview, web form), have different time periods (e.g. monthly, yearly) and different due dates. Some metrics are compatible and can be re-used while others are not. When possible we try to reuse the data, format it to our metrics schema and then ask for feedback and corrections from the responsible ESIPs. The impact of this coordination was that the total number of metrics increased. For example the DAACs had always been reporting on the volume of data delivered so we also included that metrics to allow longitudinal comparisons even though we already had a similar (and preferable) metric for the number of products delivered.

Without coordinated metrics collection, the data metrics data providers become rapidly annoyed by the multiple requests for data. Coordination is rarely achieved because it is difficult, and because it does not benefit the decision-makers who are sponsoring the data collection.

4.4 Baseline Definition

The definition of the baseline presented difficulties as our evaluation efforts began after the start of the Federation. Data related to the baseline measure before the inception of the federation is sparse and mainly anecdotal - the exception being the DAACs which existed before the start of the other ESIPs. Their performance data could be used as a basis for comparison, however, the mere number of projects grew and thus performance is expected to multiply, this comparison would not be adequate.

Many baseline items consisted of vague complaints (such as "it took too long to get data", or "too long to make new data available"). Even though no good baseline numbers are available (e.g. number of days or months), it was useful to review this

information and choose metrics or nugget categories that would help back up our claim that we were doing "better than before". As an example, the delivery time became a required metrics (average delivery time for different types of media). For the 'time to dissemination' comparing average times would be useless, we thought of counting the number of complaints - but formal complaints are rare, so instead we added a category of nuggets "Examples of rapid dissemination of new data" in which ESIPs are encouraged to report success in this direction.

In summary, even though the baseline is hard to define and not quantitative, attempting to describe it is very useful in formalizing goals and selecting metrics.

4.5 Coercing Participation in the Data Collection

The GPRA makes evaluation and metrics collection mandatory but how much effort is to be dedicated to it will remain a matter of interpretation. Some ESIPs have more resources than others and will be more likely to have staff and resources for generating good metrics data. ESIPs with poor results may not report, which might muddy the metrics summary, but those ESIPs will suffer the consequences later on when renewal time comes, so in the long term this problem should recede.

We found that merely asking the ESIPs to submit data would not be successful without providing a way for the ESIPs to also view their own data and the data of other ESIPs. Therefore we are working on password-protected data viewers that let federation participants review the collected data.

On the other hand, it was suggested that the qualitative data, such as nuggets, would be better reported globally, i.e. at the federation level, to avoid the bias favoring large ESIPs who would have more resources and could assign more staff time to enter nuggets.

4.6 Internal vs. External Evaluation

The current metrics were developed and defined with the evaluands, the organization being evaluated. These metrics are reported to the management of the project, which presents a contradiction of interests. The projects are not likely to suggest metrics that will present a negative image of their performance, as this would be against their best interests. An external evaluation body ('judge') would need to be appointed in order to insure impartiality.

4.7 Evaluation as an agent of change - You become what you measure!

Metrics collection can effect the phenomena itself. For example asking for the volume of data delivered encourages ESIPs to create large chunks of inseparable data to boost this metrics. But having files too large to download was a complaint identified in the baseline, so instead it seemed important to measure the number of items delivered, which would encourage smaller chunks so that users can download only what they need. It seemed important to favor metrics that "if abused", would have a positive overall effect on Federation activities. (note that we ended up using both metrics to remain compatible with another collection effort -see 4.3).

The best example of evaluation as an agent of change might be the high level of collaboration among the ESIPs. This can be in part attributed to the initial identification of collaboration as one of the important factors in the evaluation of the Federation by NASA.

5. Conclusion

We have described some of the challenges facing the ESIP Federation in organizing its own evaluation, and discussed the issues encountered during the metrics definition phase of the early data collection. This is just the beginning. We are currently collecting the first round of data and feedback on the appropriateness of the proposed metrics via personal telephone interviews. Tools are being implemented to allow online data collection and review. We know that this early evaluation will only provide a crude overview of the early activity of the Federation but will also provide a crucial tool to its vitality.

Our experience indicates that a large number of metrics will be needed to fully represent the activities and strengths of all partners, but because of the heterogeneity of the ESIPs the qualitative data (comments accompanying the metric data and success stories) becomes the most useful data. Other lessons learned include the need for evaluation data viewing tools to accompany collection tools. Finally, our experience confirms the effect of evaluation as an agent of change.

6. Acknowledgements

This work is supported in part by NASA (NCC5300). We want to thank all the Federation members that helped define the evaluation challenges presented here, Joseph Ja'Ja' and John Townshend, principal investigators of the Maryland ESIP, Ben Shneiderman and Gary Marchionini for their helpful feedback.

7. References

- [1] GPRA 93: Government Performance Results Act of 1993
<http://www.whitehouse.gov/OMB/mgmt-gpra/gplaw2m.html>
- [2] Marchionini, G. Evaluating Digital Libraries: A Longitudinal and Multifaceted View. Technical Report, June 2000.
<http://ils.unc.edu/~march/perseus/lib-trends-findraft.pdf>
- [3] Lancaster, F. W., If you want to evaluate your library. Imprint, Champaign, IL : *University of Illinois, Graduate School of Library and Information Science*, 1988.
- [4] Griffiths, I-M. A manual on the evaluation of information centers and services by Jose-Marie Griffiths and Donald W. King. New York : AIAA, 1991.
- [5] Marchionini, G., Crane, H., Evaluating hypermedia and learning: Methods and results from the Perseus project. *ACM Transactions on Information Systems*, vol. 12, 1 (Jan. 1994) 5-34. also: Evaluation of the Perseus Hypermedia Corpus.
<http://www.perseus.tufts.edu/FIPSE/report-final.html>
- [6] Hill, L. et al. (1997) User Evaluation: Summary of the Methodologies and Results for the Alexandria Digital Library, University of California, Santa Barbara. In: *ASIS '97. The Annual Meeting of the American Association for Information Science*.
<http://www.asis.org/annual-97/alexia.htm>)
- [7] National Research Council. Computer Science and Telecommunications Board. Design and Evaluation: A Review of the State of the Art. *D-Lib Magazine*. July/August 1998.
<http://www.dlib.org/dlib/july98/nrc/07nrc.html>
- [8] A. Bishop, B. Battenfield, & N. VanHouse (Eds.) *Digital library use: Social practice in design and evaluation*. MIT Press, Cambridge, MA (2000)

8. Related Sites

Federation website: <http://www.esipfed.org/>

Federation evaluation resources and metrics documents website:

<http://esip.umiacs.umd.edu/documents/eval/>