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

Catherine Plaisant, Anita Komlodi[#]

Human-Computer Interaction Laboratory
University of Maryland Institute for Advanced Computer Studies
[#]College of Information Studies

University of Maryland, College Park, MD 20742
TEL (301) 405-2768 - FAX (301) 405-6707
plaisant@cs.umd.edu, komlodi@glue.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 cheaper access to an expanding earth science user community. As it is organizing itself, the federation is mandated to set in place an evaluation methodology and collect metrics reflecting the health 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.

Keyword: metrics, quantitative evaluation, qualitative evaluation, earth science

1 Introduction

Beside the obvious need to evaluate any experiment to measure its positive and negative impact the impact of the Government Performance and Results Act (GPRA) 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 Government Performance and Results Act (GPRA) of 1993 was created as a 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', 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. Output and outcome measures, performance goals and indicators are defined.

NASA and the many programs it sponsors also promote evaluation in every aspect of the work they support. In this paper we will 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 unique 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 an ESIP Federation could be the viable enterprise model to facilitate the public availability of data from geographically dispersed providers. The Federation itself is charged with developing this new enterprise model. It is charged to define a governance method that encourages cooperation as well as create a community of ESIPs that are dedicated to the continued success of the Federation.

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

Evaluation is a very practical field, it is intended to improve a program or system. It applies research methods to assess the value of the object of the evaluation. It may serve the purposes of decision making, help selection among different alternatives. Lancester (1993) suggests defining inputs, outputs, and outcomes in evaluation. The long-term objective of any information 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.

The object of evaluation can be many different entities, Owen describes the different classes of 'evaluands': planning, programs, policies, organizations, products, and individuals. The current evaluation is twofold, it evaluates the Federation as an

organization. The more important goal is to evaluate it as a program, and as a service or product. This multiplicity of goals added to the complexity of the problem.

Information system and service evaluation can be used as a framework for the evaluation of web-based information services, as the major operations are very similar: collection, processing, organization, and provision. 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.

The technology is different, and this has implications for evaluation. Interface and software performance evaluations need to be integrated into the process to cater to this aspect of the program. Users in the case of web-based systems are geographically more dispersed, and assessing their characteristics and needs may present logistic problems. Libraries are old institutions, users go through library education in schools and have reasonable expectations and assumptions about the operation and services of libraries.

Lancester suggests examining inputs, outputs, and outcomes in this order, which also represents increasing complexity. Outcomes are often long-term and related to behaviors that are hard to measure. On the other hand, inputs and outputs are easily quantifiable. Outputs must be evaluated in terms of quality as well. "Criteria used to evaluate outputs should be good predictors of the extent to which the desired outcomes are achieved.

Marchionini (1994) provides one of the best example of long term digital library evaluation. The author describes methodology, results and implications. 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.

In the domain of geographical information systems Hill (1997) 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, and in Bishop et al (2000).

3 The Earth Science Information Partnership Federation

One challenge for the ESIP Federation is the prototyping of alternative ways of developing, producing, and distributing environmental 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 emerging discipline of Earth System Science. Earth Science research currently under development has the potential to yield a variety of new scientific understandings and practical benefits, from understanding monitoring deforestation or understanding global changes to providing customized reports to farmers

or fishermen, or services to local government and land owners. The ESIP Federation can provide the means to manage, and develop these efforts.

The ESIP federation was launched in 1998 with the selection by competition from the government, academic, and private sectors of the initial ESIPs.

The most striking characteristic of the Federation might be the diversity of its partners:

- 10 "Veteran" Data Access and Archive Centers (DAACs) who have been in place for years and deal with the archiving and distribution of data (called ESIP1s).
- 12 new data providing ESIPs, generally run by scientists themselves at Universities create products for the research community (called ESIP2s).
- 12 commercial entities, from startup companies to established educational entities like museums, provide practical applications of earth science data for a broader community (called ESIP3s – they are only partially funded by NASA, and meant to become rapidly self sustaining)

ESIPs have different levels of funding and cost sharing, they are at different stages of developments – from well established to not quite public yet, some offer data only, other services only, some both. Some ESIPs have a handful of employees, other several dozens. The Federation will be evaluated by the level of cooperation among ESIPs, but individual ESIPs were selected following a competitive process and may have to compete against each others 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 definition.

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

4 Challenges of Evaluation

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 program development purposes for their own use, collect data for management for different purposes.

The development of evaluation metrics across all units will inevitably involve representation from these units. 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 will be decided based on management decision or voluntary selection. Our 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 different sets of metrics. We defined a new set based on these and our evaluation experience with the Global Land Cover Facility (one of the ESIP 2 projects at the University of Maryland http://glcf.umiacs.umd.edu).

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:

- limiting the number of metrics
- identifying metrics that truly represent the activities and strengths of all entities

In order to accommodate this variety, we chose to collect both quantitative data (metrics) and qualitative data (in the form of nuggets or success stories)

Several approaches emerged for the metrics. A first approach involved defining a small, common set of mandatory metrics that all projects would provide, while creating a larger group of optional metrics. This also included a third level of metrics, custom metrics defined and collected by the individual ESIP in order to improve their own activities (e.g. what part of the website is most used, is the help system being used).
This approach was rejected due to the lack of agreement on a small common set of metrics. Instead, a larger set of common metrics were developed, acknowledging that not all ESIPs will be able to collect all metrics and would just mark them as Non Applicable.

The other direction chosen to insure flexibility and satisfy all ESIPs was to collect data in the form of 'nuggets'. These are success stories, short descriptions of data reporting on characteristic activities and accomplishments of the different projects. There are currently twenty categories of nuggets defined, based on discussions with the different projects. This allows us to tabulate the number of nuggets in each categories.

Examples: the initial list of metrics included measuring the amount of data delivered, the volume of data etc. because this was the traditional way of measuring EOSDIS data centers. But some of the ESIP do not have any data per se, but provide valuable services processing data provided by users or services using third party data (e.g. for legal services, or museum applications), so we had to also measure the number and type of services. Original metrics looked for total number of users, so we were adding 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 - generally unidentified) from data and service users (who download or order data and are identified).

Even after adding many metrics to our list (many being non-applicable for any given ESIP) there was still a strong feeling of not being able to completely picture the richness of activities. For example one of the ESIP didn't have much data, very few users but generate results that could be evaluated by the number of lives they saved! Clearly not a

metric applicable to many ESIPs. This case was handled by creating a category of qualitative metrics for reporting this kind of data.

Examples of metrics:

Number_of_data_and_service_users
Categorization of data_and_service_users by domain
Categorization of data_and_service_users by market share
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

Examples of nugget types

New science:
- New type of data use
- New types of data products now available to the community
- Data quality achievement

Federation activities
- Notable results from working group or committee
- Example of federation collaboration
- Collaboration with other institutions

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

Dissemination:
- Publications written by ESIP users referring to ESIP data
- PR events (e.g. Presentation by Esip staff at conferences or community events)
- Mention of ESIP in press

Education
- New K-12 education activities generated from ESIP
- New higher education or professional education activities generated from ESIP
- Student graduated whose work was directly linked to ESIP

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

Baseline Definition

The definition of the baseline presented difficulties, as our evaluation efforts began well after the start of the projects. Data about the situation before the existence of the federation is parse and mainly anecdotal. Nine members of the Federation existed with

similar missions 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" e.g. it took too long to get data, and 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 to choose metrics that would back up our claim that we were doing "better than before". For the delivery time, it became a required metrics (average delivery time for different types of media). For the time to dissemination each case is a special case so it would be unfair to ask for average numbers, so 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 if the baseline is not clearly defined and usually not quantitative, attempting to describe the baseline is nevertheless very useful to define goals and select metrics.

Coercing Participation in the Data Collection

GPRA makes evaluation and metrics collection mandatory but how much effort is to be dedicated to it will always be 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. 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 will recede.

We felt strongly than merely asking the ESIPs to submit data would not be successful if we did not provide a way for the ESIPs to also see their own data and the data for other ESIPs. In parallel to the development of online data collection forms, we are also working on password protested data browsers that lets federation participants review the collected data and compare their activities with those of other ESIPs.

On the other hand, the was a strong feeling that the more anecdotal 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 more nuggets.

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’) will need to be appointed in order to insure impartiality.

Evaluation as an agent of change

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 is a known 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 could only download what they need. But this could also have an adverse effect, we decided to count both volume and number.

In summary it seemed important to favor metrics that "if abused", would have a positive overall effect on Federation activities.

Dealing with past or parallel collection efforts

Because of heterogeneity of the ESIPs subsets of ESIPs are also part of other metrics collection efforts. For examples the DAACs (ESIP1s) have been in existence for a while and some 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 are most likely to use different collection mechanisms (interview, web form, etc.), have different time periods (e.g. monthly or yearly) and start dates. Some metrics are compatible and could potentially be re-used while others are not. When possible we plan to reuse the data, convert it to our metrics and then ask feedback and corrections from the concerned ESIPs.

The challenge here is that without a centralized or highly coordinated metrics collection, the entities having to provide those metrics become rapidly annoyed by the uncoordinated requests. But this coordination is hard to achieve because this coordination mainly benefits the metrics data providers, not the evaluators or the decision makers who are sponsoring the parallel data collection efforts.

5 Conclusion

We have described the challenges facing the ESIP Federation in the organizing of its own evaluation, and discuss the issues encountered during the metrics definition phase of the early data collection. This is just a beginning. We are currently collecting a first round of data via personal telephone interviews and follow up to collect estimates and feedback on the appropriateness of the proposed metrics. 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 vital tool to testify to its vitality.

Acknowledgements
This work is supported in part by NASA. We want to thank all the ESIP members that helped define the evaluation challenges presented here and in particular Frank Lindsay for his contribution to this work.

References

A. Bishop, B. Buttenfield, & N. VanHouse (Eds.) Digital library use: Social practice in design and evaluation. MIT Press, Cambridge, MA (2000)

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)

Lancaster, F. W., If you want to evaluate your library. Imprint, Champaign, IL : University of Illinois, Graduate School of Library and Information Science, 1988.

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)

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)

Related Sites

Federation website: http://www.esipfed.org/
Federation evaluation resources and metrics documents website:
http://esip.umiacs.umd.edu/documents/eval/