A Linked Fusion of Things, Services, and Data to Support a Collaborative Data Management Facility

EG Stephan, TO Elsethagen, AS Wynne, C Sivaraman, MC Macduff, LK Berg, WJ Shaw
(eric.stephan, todd.elsethagen, adam.wynne, chitra.sivaraman, matt.macduff, will.shaw, larry.berg)@pnnl.gov
Pacific Northwest National Laboratory

***Abstract*— The purpose of this paper is to illustrate the use of semantic technologies and approaches to seamlessly link things, services, and data in the proposed design of a scientific offshore wind energy research for the U.S. Department of Energy Wind and Water Technology Office of the Office of Energy Efficiency and Renewable Energy (EERE). By adapting linked community best practices, we were able to design a collaborative facility supporting both operational staff and end users that incorporates off-the-shelf components and overcome traditional barriers between devices, resulting data, and processing services. This was made largely possible through complementary advances in the Internet of Things (IoT), semantic web, Linked Services, and Linked Data communities, which provide the foundation for our design.**

Keywords - Internet of Things, Linked Data, semantic web, linked services, atmosphere, data management facility

I. INTRODUCTION

Advances in linked communities have captured the imagination of the applied scientific community by ushering in the beginning of the golden age of seamless and open integration between the essential components of scientific research: things, knowledge, data, and services. *Linked fusion* is an informal term we use to describe the complementary and cooperative approaches being introduced to show how linked technology can be fused together to solve complex scientific experimental problems.

Throughout this paper, we discuss *things*, *services*, *data*, and *knowledge* from a scientific research project perspective linking these seemingly disparate items together into a project-specific linked web. *Things* we informally define as any electronic device or non-electronic infrastructure that the electronic device is located on or nearby that could impact the electronic device (e.g. *in situ*, platform, boom etc.). *Services* we generically define as automated machine processes that handle data streams or data files as well as tracked human activities that may impact things or data. *Data* we define as physical file containers and directory trees representing a particular data set. *Knowledge* we define as any metadata, thesauri, or RDF vocabulary used to describe syntactic data structures or semantically represent things, services, and data.

A decade ago, from an application developer perspective, the world of linked technologies seemed to be separated into dramatically different focus areas. The Internet of Things (IoT) community focused on innovating approaches to connect devices to the internet. The semantic web community focused solely on portraying knowledge and data as a web of interrelated terms and concepts. In a similar fashion, the World Wide Web community sought after ways to link browsable information together. Web-based services communities in like manner enabled business enterprises to develop service-oriented approaches to solving business problems. When Representational State Transfer (REST) design principles [1] emerged, web API communities began calling anything of perceived importance a resource (e.g. data, knowledge, and services). These resources are addressable by a meaningful name, Uniform Resource Identifier (URI), and capable of being linked to other resources [2]. By handling data, information, and services as resources that could be linked together in any fashion, the pendulum swung from requiring application developers to be service-oriented, thing-oriented, or data-oriented to more of a complementary approach linking any type of resource together despite the disparity. Linked Data had design goals similar to REST, making data addressable, accessible, and interconnected with knowledge [3]. This Linked Data community led to domain-specific applications that provided functionality through *data mash-ups*. Data mash-ups are simply strategies to incorporate data from various linked data sources to solve a domain-centric problem [4].

Table I: LINKED DATA 5 STAR DEPLOYMENT SCHEME[5]

Rating	Criteria
★	Data is available on the Web, in whatever format.
★★	Available as machine-readable structured data, (i.e., not a scanned image).
★★★	Available in a non-proprietary format, (i.e. CSV, not Microsoft Excel).
★★★★	Published using open standards from the W3C (RDF and SPARQL).
★★★★★	All of the above and links to other Linked Open Data.

Following the advances in the Linked Data community, the Linked Services concept was developed as the next wave of service technology based on two ideas: make service annotations available in the web of data, and create services to

COLLABORATECOM 2013, October 20-23, Austin, United States

DOI 10.4108/icst.collaboratecom.2013.254158

support the web of data [6]. Linked services made services addressable, accessible, and self-describing with knowledge using a common RDF vocabulary that generalized the concept of service, and acknowledged varying technical implementations. Recently, the Linked Services community made the connection to the IoT community [7] referencing things like services using the same RDF vocabulary.

Due to these advances, and given the level of maturity in these research efforts, we contend that scientific applications should seriously consider leveraging these linked design principles to support their scientific end-to-end solutions, particularly in cases where things, services, and data need to be managed.

II. USE CASE

This use case illustrates how we plan to implement *linked* best practices in a real world problem space involving the integration of things, services, data, and knowledge to support quality scientific research.

The intent of the Reference Facility for Offshore Renewable Energy (RFORE) Data Management Facility (DMF) is to make the planned instrument measurements taken at an off shore reference facility accessible to a range of users, which includes atmospheric researchers and energy planners who perform wind energy studies. It can also be used to disseminate a portion of near real time measurements to weather forecasters. To achieve the scientific research goals, the DMF needs to provide:

- Reference data for validating new measurement technologies for resource characterization and assessment for offshore renewable energy
- Suitable data to support research that fills important knowledge gaps in the characterization, assessment, and forecasting of offshore renewable energy
- Data appropriate for ingest into weather forecast models to improve local renewable energy resource forecasting.

These functional requirements identify the need to receive, manage, preserve, and make high-integrity measurement data both easily searchable and accessible. Because the RFORE project study time period is being performed for an indefinite period of time, data management and preservation are being treated separately to support both day-to-day operational activities as well as long term archival that includes information preservation so data can be easily recalled with contextual relevant information about the measurements as they are handled by future users.

A. Facility Interfaces

1) Producers

The reference facility is considered the main producer of data in the form of instrument measurements. Every five minutes, instrument data will be automatically transmitted to the DMF. For operators and instrument mentors who monitor and maintain the instruments, the DMF will provide the means to log information. This information includes: problem reports, maintenance visits, instrument calibrations, and quality assurance reports.

2) Consumers

Consumer communities, also known as users, are split into four main categories: RFORE operations, weather forecasting, energy and weather profiling studies, and ad-hoc searches. Operations users support any activity relating to RFORE operations, instrument mentoring, or quality assurance. They will access the DMF on a frequent basis to support day-to-day activities. Weather forecasters will have access to monitoring tools and data. Some data, once it arrives in the DMF, will be pushed to forecasting agencies. Sciences focusing on energy and weather profiling studies will perform query-by-example (QBE) searches, faceted searches, or advanced searches to find and download measurements based on their study criteria. Finally, for advanced users, the DMF will support ad-hoc searches for very specific complex questions.

3) Management

Management includes all DMF staff responsible for administering the facility day-to-day operations, including reporting any problems detected in data transfers and any DMF functions. Management also includes monitoring processes responsible for reporting or responding to unplanned outages and faults on any aspect of the data pipeline.

III. RFORE LINKED DESIGN PRACTICES

A key challenge on RFORE is strategically leveraging state-of-the-art, off-the-shelf standards and technology with a proven track record in operational data management facility support.

A primary contributor of standards and technology from the atmosphere data management perspective is the Atmosphere Radiation Measurement (ARM) DMF team (http://www.arm.gov). Throughout the planning process, the RFORE team actively collaborated with the ARM team to identify software and data standards that will be leveraged by the RFORE DMF.

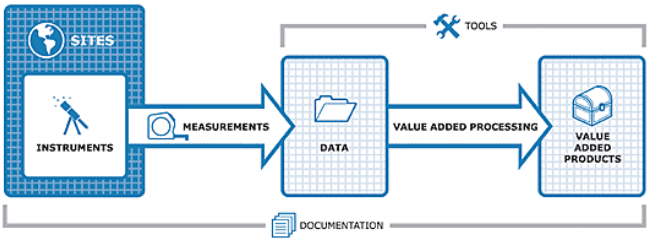

Figure 1: RFORE was able to leverage many capabilities from the ARM DMF operational facility [8]

Table II depicts some capabilities required by the RFORE DMF and corresponding identified technology. Despite the diverse array of off-the-shelf technology, the goal of the RFORE DMF is to provide users with rich seamless access to their data. For example, users examining temperature measurement data retrieved from the DMF archive may need to assess any possible inaccuracies associated with the instrument.

To properly assess any inaccuracies, users will need access to historical records related to the instrument. Examples of historical records may include instrument problem reports, quality assessments, and maintenance reports all contained in the Activity Tracking system. From the user perspective, they may not know the specific instrument that took the measurements, but nonetheless they would still need to have a catalog correlating instruments, measurement types, and archived data.

TABLE II: RFORE DMF INVENTORY OF OFF THE SHELF COMPONENTS REQUIRING INTEGRATION

DMF Capability	**Off the Shelf Provider**	**Interfaces**
Catalogs	open source, government	RDF
Collection	government	File API
Ingest	government	REST
Processing	government	REST
Archive	commercial	File API
Problem Reporting/ Quality Assurance	commercial	REST, SQL, Web
Change Management	commercial	REST, SQL, Web
Types, vocabularies, standards	open source	File APIs, semantic web
Access	open source	File access, REST, semantic web, SQL, Web

A. Common Addressible Layer

Based on the capabilities described in Table II, it was important to make sure that any resource important to RFORE (devices, services, activities, data, and metadata) were all identified with a common unique identifier strategy. With the linked design principles described earlier, we determined that every resource will have a unique and meaningful name (URI) from the DMF perspective. For some resources such as archived data, the URI will signify an addressable location, for others it will provide a unique reference to a resource in the reference facility. For all identifiers, the name will be intuitive and meaningful to the RFORE user community and part of its standards. Filenames in archived datasets will carry a standardized naming convention as well based on the ARM naming standard, as shown in Fig. 2.

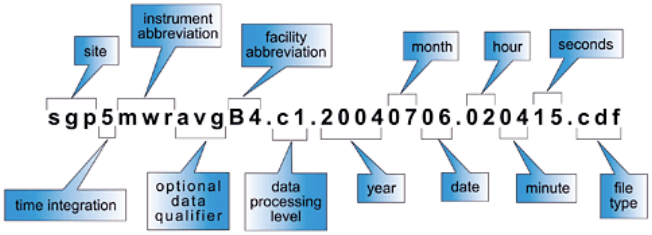

Figure 2: Example of ARM standard naming convention that will be used on RFORE [9].

B. Cataloging

Although every resource will be uniquely identifiable in a meaningful way, additional referential technical information is required to help describe each in greater detail. Catalogs are regarded as the linchpin of the linked fusion concept since they possess the logic for users and machines to search, discover, and reason semantic and syntactic information relating to things, services, and data. ProvEn Services [10] is a knowledge repository and will be used to store each catalog. Each catalog will rely on foundation vocabularies that will be manually aligned to each other to support cross-reference searching, and linked fusion between catalogs. At this time, we are anticipating at least four types of catalogs:

1. facility, instruments, and measurements
2. activities and services
3. data
4. vocabulary management

1) Facility, instruments, and measurements

This catalog is designed to provide the DMF knowledge about reference facility electronic devices and its infrastructure. Currently, the facility is targeting 25-30 different types of *in situ* and remote sensing devices collecting 42 GB of data per day. Infrastructure may include the facility platform itself, the booms on which instruments are installed or other items at the facility. While there is no direct electronic link to the facility infrastructure, the catalog will help make them "smart" by collecting monitoring information gathered from the maintenance log, scheduling systems, and change management logs. Visits to the platform will be via helipad. Knowing the arrival and departure schedules will be important because the anemometer wind measurements could be affected by each visit.

2) Activities and Services

Having referential information and versioning information about operations and instrument mentor activities, such as facility visits, quality assurance, maintenance change management, calibration problem reporting and problem resolution, is important from a data integrity perspective to provide full disclosure of work with data or reference facility devices. Services involved in the collection, transfer, ingest, and processing equally needs to be tracked.

3) Data

All archived raw measurements, processed measurements, related calibrations, etc., will be tracked in the data catalog. The RFORE ingest process will be responsible for updating newly archived collected and processed data.

4) Vocabulary Management

Based on best practices [12] and recommendations from active linked data and domain specific communities, RFORE will host a separate catalog to organize and align RDF vocabularies to one another.

C. Linked Criteria

To document how RFORE *things*, *services*, and *data*, are linked together and accessible we adapted the Linked Data 5 Star Rating System (Table I). The original Linked Data "Web" criterion was based on accessibility to anyone. The

RFORE Web criterion instead restricts its user community to operators, scientists, instrument mentors, forecasters and other collaborators who need access to RFORE resources (Fig. 3,4). Table III illustrates how example RFORE things are rated.

TABLE III: RFORE 5 STAR LINKED THINGS CRITERIA

Rating	Things	Criteria
★	All instruments	Accessible on the RFORE Web, in whatever format.
★★	All instruments Related activity tracking	Accessible as machine-readable structured data
★★★	Instrument log files, some measurements	Accessible in a non-proprietary format, (i.e, CSV, text).
★★★★	Instrument catalog	Published using community vocabularies and domain standards
★★★★★	Instrument catalog	All of the above and links to services and data

For this information to be useful, this rating information needs to be included in the catalogs mentioned in the previous section.

IV. Discussion

Given the level of complexity discussed in the use case, there were risks we felt were important to point out if linked principles were not applied.

A. Missing Addressible Layer

Instruments, databases, services, and activity tracking systems come with a default way of identifying resources. Without a governing naming policy resource identifier risk being unstandardized, not intuitive to humans, and not easily integrated by software.

B. No Cataloging

Without the means to create an overarching layer of knowledge semantically linking things, services, and data, we felt the DMF might suffer from several problems: information silos, difficulty in data/information exchanges, and a lack of comprehensive provenance tracking.

1) Information silos

If 3rd party systems are used in an operations setting without an integration stragegy, information silos can result. These silos tend to limit the productivity of an organization because of the cost/time it takes to informally exchange information between systems.

2) Difficulty with data/information exchanges

This is somewhat related to the problem of the previous discussion of information silos. Without the use of data standards, data analysis for the end users might be extremely haphazard leading to misinterpretations of data, the purposes of services, and how the different components relate to each other.

3) Lack of comprehensive provenance tracking

This issue might not be particularly apparent in the initial implementation of the DMF, but the lack of an overall plan to capture and convey the origin of data and subsequent modifications could lead to unanswered questions. Because RFORE collected data needs to be stored indefinitely, provenance information describing the origin of the measurement data also needs to be retained. Provenance [11] is used to provide consumers of data a complete understanding of how the data was collected, the history of the instrument that measured the data, what was done when it was originally processed, and whether any transformations took place on the data set.

C. Missing Linked Criteria

It would be naïve to simply choose a linked approach without some way of planning and documenting how things, data, and services are linked together. Without a clear set of criteria for linked resources, the linked solution risks being inconsistent and difficult to navigate.

D. Additional Caveats For Consideration

Choosing to use a linked strategy in an operational research facility comes with some considerations that need to be weighed carefully. While other system integration options have not been the focal point of this paper, we have in essence shifted the integration problem from a software maintenance realm to a knowledge management realm. Similar to software development configuration management practices, procedures are needed to identify a clear concept of operations for managing knowledge. Policies will need to be established for updating new versions of vocabulary while supporting backward compatibility for currently used vocabularies. A clear concept of operations needs to be in place to maintain all catalogs as devices, services, and data are updated and changed. To automate this process, the ingest process, and activity tracking components will update and maintain the catalogs on a routine basis.

V. Conclusion

Because of the cooperative and complementary approaches of linked communities we were able to adopt *Linked fusion* to design a facility capable of linking things, services, and data together with knowledge to support complex scientific experimental problems such as the RFORE data management facility. Using linked design practices, we were able to provide an overarching approach to linking disparate systems together and avoid the pitfalls of providing solutions that might either be difficult to maintain or provide users with a

segregated solution that might be difficult for them to navigate.

VI. References

[1] R Fielding, Roy Thomas. Architectural styles and the design of network-based software architectures. Diss. University of California, 2000.

[2] Richardson, Leonard, and Sam Ruby. RESTful web services. O'Reilly, 2008. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] Heath, Tom, and Christian Bizer. "Linked data: Evolving the web into a global data space." Synthesis lectures on the semantic web: theory and technology 1.1 (2011): 1-136.

[4] Berners-Lee, Tim. "Linked data-the story so far." International Journal on Semantic Web and Information Systems 5.3 (2009): 1-22.

[5] B Berners-Lee, T. (2010). Open, Linked Data for a Global Community. In Gov 2.0 Expo, Washington, DC, May 25-27, 2010.

[6] B Pedrinaci, Carlos, and John Domingue. "Toward the Next Wave of Services: Linked Services for the Web of Data." J. ucs 16.13 (2010): 1694-1719.

[7] Mandler, Benny, Fabio Antonelli, Robert Kleinfeld, Carlos Pedrinaci, David Carrera, Alessio Gugliotta, Daniel Schreckling et al. "COMPOSE--A Journey from the Internet of Things to the Internet of Services." In Advanced Information Networking and Applications Workshops (WAINA), 2013 27th International Conference on, pp. 1217-1222. IEEE, 2013.

[8] Gaustad, Krista, et al. "The development of QC standards for ARM data products." Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing. ACM, 2010.

[9] ARM. Data Management and Documentation Plan. http://www.arm.gov/data/docs/plan. September 10, 2013.

[10] Stephan EG, TO Elsethagen, K Kleese van Dam, and LD Riihimaki. 2013. "What Comes First, the OWL or the Bean? Creating Reusable Scientific Software with OWL/RDF Vocabularies." In First Workshop on Sustainable Software for Science: Practice and Experiences. (submitted)

[11] Stephan EG, P Pinheiro da Silva, and K Kleese van Dam. 2013. "Bridging the Gap between Scientific Data Producers and Consumers: A Provenance Approach." In Data Intensive Science. Chapman and Hall/CRC, Boca Raton, FL.

[12] Hebeler, John, et al. Semantic web programming. John Wiley & Sons, 2011

VII. FIGURES

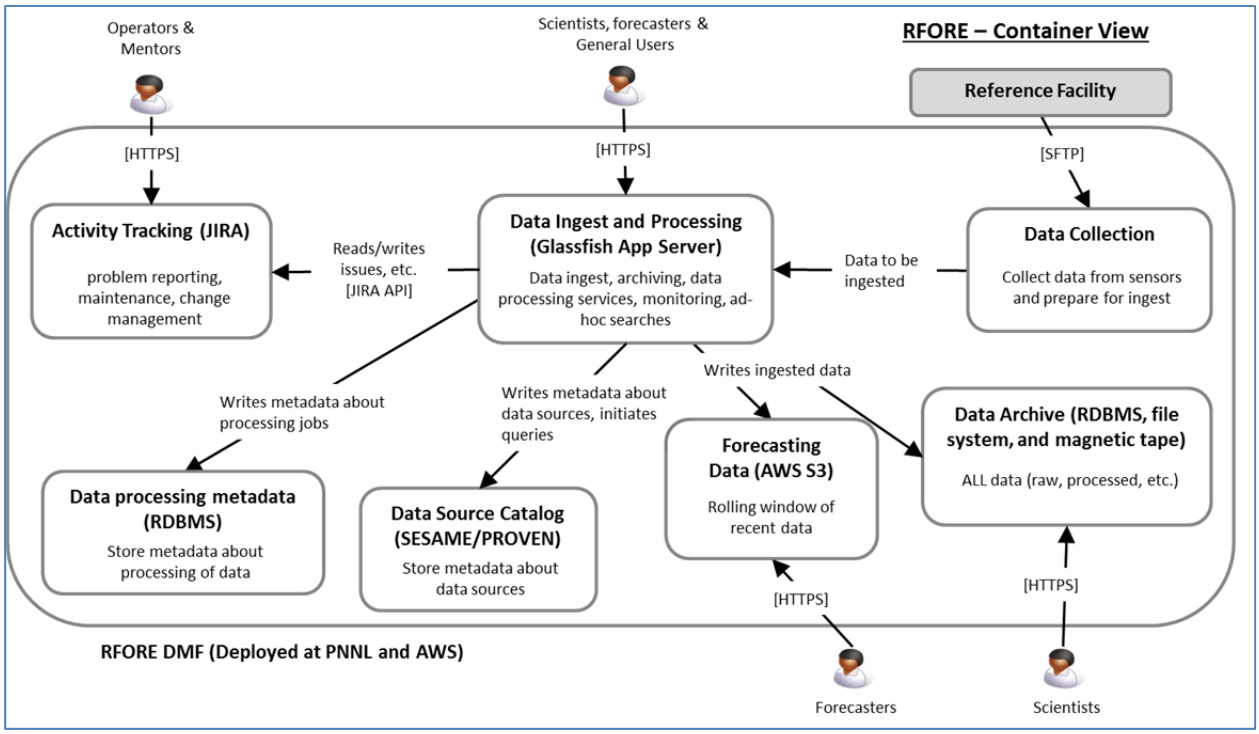

Figure 3: Conceptual RFORE DMF architecture

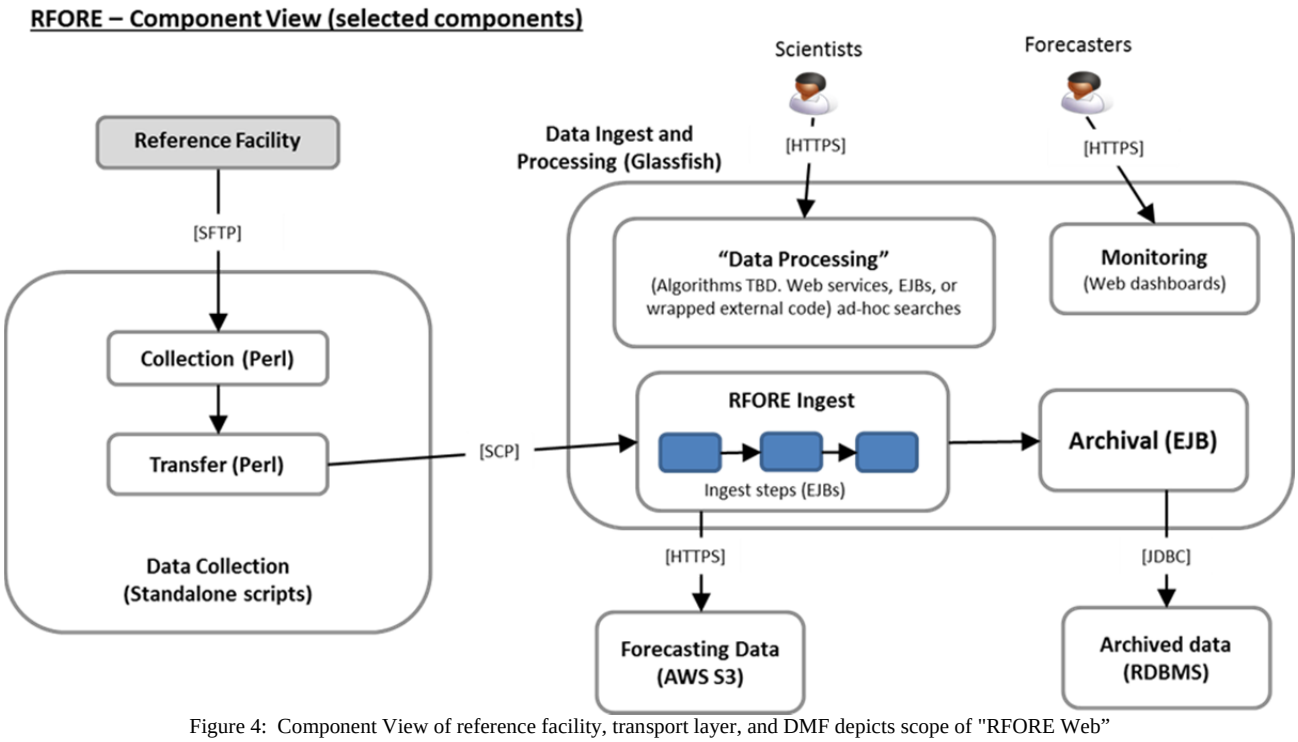

Figure 4: Component View of reference facility, transport layer, and DMF depicts scope of "RFORE Web"