The Architecture and Design of a Community-based Cloud Platform for Curating Big Data

Sulayman K. Sowe
Information Services Platform laboratory,
Universal Communication Research Institute, NICT.
3-5 Hikaridai, Seika-cho, Kyoto, 619-0289, Japan.
sowe@nict.go.jp

Koji Zettsu
Information Services Platform laboratory,
Universal Communication Research Institute, NICT.
3-5 Hikaridai, Seika-cho,Kyoto, 619-0289, Japan.
zettsu@nict.go.jp

***Abstract*—The digital universe is exponentially producing unprecedented volume of data that has brought benefits, as well as fundamental challenges for the research community. This trend is inherently exciting for the development and deployment of cloud platforms to support research communities curate data. The excitements stems from the fact that researchers can now have more access and discover more value in data, establish relationships between bits and pieces of information from many types of data, and collaborate with a diverse community of researchers from various domains. Technically, however, platform providers must design their infrastructure in such a way that users can easily search, find, share, and seamlessly collaborate. In this paper, we present the architecture and design of a cloud platform and describe how a community of disaster response scientists could use the platform to curate data. Motivation for developing the platform, lessons learnt in overcoming some challenges associated with collaborative and cooperative cloud environments, and future research directions are also presented. The contribution of this research is in the field of cloud infrastructure for supporting data curation activities, and the sustenance of research communities for the utilization of information assets and cyber technologies.**

***Keywords*-Cloud computing; Big data; Data curation; Scientific communities; Collaborative environments.**

I. INTRODUCTION

The exponential growth of big data represents a significant trend in the cloud computing and data-intensive science era [1], [2]. This trend is likely to persist, given the fact that man and machines continues to produce large volumes of data at an unprecedented rate. For example, the IDC study [3] predicts that by 2020, the digital universe will house 40,000 Exabytes of data. Paradoxically, the investment per gigabyte will drop from $2.00 to $0.20, by the same period.

Big data is characterized by what has come to be known as the 6Vs (Volume, Variety, Velocity, Veracity, Visualization, Value), and technology pundits are still debating what is and what is not big data. According to [4], big data is enormous amounts of unstructured, sometimes structured or semi-structured data produced by high-performance applications falling in a wide and heterogeneous scenarios. However, [5] argued that big data is "less about data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets". In the authors' view, the value of big data comes from the patterns that can be derived by making connections between pieces of data, about an individual, about individuals in relation to others, about groups of people, or simply about the structure of information itself.

We posit that it is this "*techo-socio-network*" or the ability to support interconnections between applications, data, and people that make big data so appealing to the research community. Just like the "Repository of Repositories" concept [6], with big data, researchers can have access to more data, establish more relationships between various types of datasets from diverse sources and domains, and collaborate with other big data researchers.

There is ample anecdotal and empirical evidence detailing the impacts of this big data tsunami for research communities [1], [2], [7]; disaster response [8], [9]; economic and business [8], [10]; medicine and healthcare [11]; genomics [12]; Open (government) data; ubiquitous analytics [13]; and the development and sustenance of Cyberinfrastructure [7], [14]. Researchers in these domains highlighted big data opportunities, privacy, policy, and security issues, and technical challenges such as computational and analysis techniques, linking and storing large datasets, and platform requirements.

A. Research Questions and Contributions

To date, little research attention is focused on the people and communities who are the real beneficiaries, and, at the same time, generators of data streams that contribute to big data. Less we understand how people collaboratively create big data information assets, what their motivations may be, what challenges they face in generating, archiving, and curating big data. However, lessons from computer-supported cooperative work [1], [9], [15] and knowledge sharing practices in Open Source software development [16] shows that coordinating the activities of (big data) communities across space and time can be problematic. Furthermore, there is diminutive (if any) best-practice guidelines, design patterns, requirements, use cases, etc. to help us understand

978-0-7695-5106-7/13 $26.00 © 2013 IEEE
DOI 10.1109/CyberC.2013.35

what the characteristics are or know how should a cloud platform that support research communities curate big data look like. Taking these research gaps into consideration, the primary aim of our research is to answering the following question:

> **Research Question**: *How can we design, develop, and maintain a cloud platform that can support scientific communities curate big data?*

To help use answer this research question, we first describe the architecture of the cloud platform for curating big data we are developing at the Information Services Platform (ISP) lab of the National Institute of Information and Communications Technology (NICT), Japan. Then, we discuss how scientific communities such as disaster response scientists are using the platform to curate big data.

The architecture employs standard cloud computing models, but has extensive customized applications that we believe are vital for supporting collaborative development of big data information assets. In describing the disaster response community, we introduce the data curation model that shows how researchers check-in and check-out data from an information assets repository, and continuously get involved in data curation activities. The model encourages collaboration between data providers, consumers, and curators.

We hope that this exposition will contribute to the R&D debate surrounding cyber-related technologies, with a specific focus on the architecture, models, and application of scientific computing in the cloud. The data curation activities has the potential to increase our understanding about the community dynamics in an interconnected, distributed, and data-intensive research environment. Furthermore, we posit that this research could offer some useful insights for researchers, manages, and (cloud and big data) infrastructure and applications developers who are providing ubiquitous services for their users.

The rest of the paper is organized as follows: In section II, we discuss the background, related work, and clarify concepts we used in our research. In section III, we describe the architecture of the cloud platform we have developed to support disaster response scientists curate big data information assets. We introduce the disaster response community in section IV, and describe use cases that enable community members take full advantage of the platform functionality. The data curation model that depicts collaborative development of information assets by the disaster response community is presented in section V. In section VI, we present the Web portal of the disaster response community, and briefly discuss the platform and community sustainability measures we are undertaking, as well as research limitations. We conclude our research in section VII.

II. Background and Related Work

The dawn of cloud computing and big data has produced a fundamental shift in the way data is being generated, managed, and utilized by research communities. Scientist in all fields are increasingly querying and analyzing large datasets to help them make informed choices, discoveries, and bring about innovation. However, despite advances in technological infrastructure to support workflow management and collaboration and cooperation, the management of data, including its preservation, and the coordination of research communities remains as challenging as ever. Some of the challenges we face includes diversity both in terms of data and applications for processing and analysing data; widening gap between researchers and data providers; organizing and managing multiple and changing roles of stakeholders (R&D institutions, research labs, companies, governments, etc.); data security and privacy policies; and open standards to deal with the big data influx.

Furthermore, for the research community, the situation is becoming rather too intriguing. For instance, even analyzing and interpreting a "simple" earthquake situation can affect many fields. For example, a researcher might need to use historic archives of past earthquakes for comparative purpose, access satellite and geospatial data to map and visualize the affected region, weather and climate data such as rain/wind forecasts, mobile data from users in the affected region, traffic and medical data for the management of relief and evacuation, personal records for the identification of victims, government policy, economic data to count the cost of damage and recovery operations, etc.

Part of the solution to these challenges might be for scientific communities to consider adopting "good" data or digital curation practices. In broad terms, [17] defined digital curation as involving "maintaining, preserving and adding value to digital research data throughout its lifecycle". Data curation, according to [18], can be defined as "a means to collect, organize, validate, and preserve data". Lord, et al. [19] defined data curation as "the activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use". Nevertheless, Karen and Baker [20] recognized that data curation is not a static process, but a set of repeatable activities focusing on tending data and creating data products within a particular arena, such as disaster response. Data curation profiles and toolkits [21] have been employed in many disciplines to help scientists curate data.

A data curation profile (DCP), shown in Figure 1, can be seen as an "instrument that can be used to provide detailed information on particular data forms" [21]. As such, a researcher's DCP documents all the actions that might be needed to ensure full utilization and sharing of a particular dataset used in a given research context.

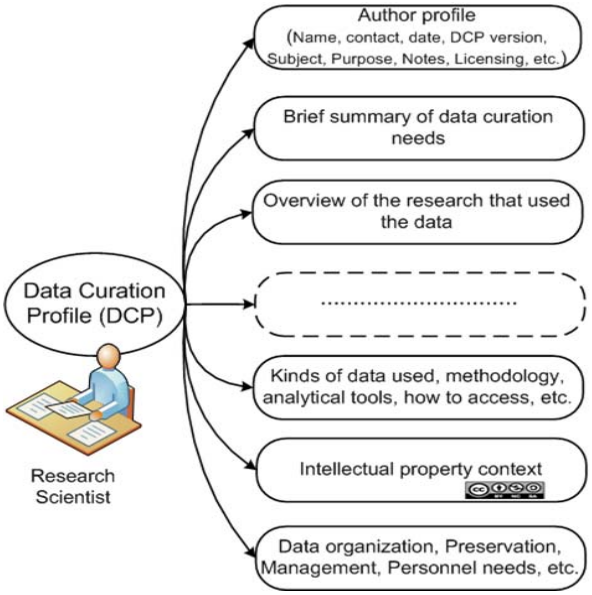

Author profile
(Name, contact, date, DCP version,
Subject, Purpose, Notes, Licensing, etc.)
Brief summary of data curation needs
Overview of the research that used the data
....
Data Curation Profile (DCP)
Kinds of data used, methodology, analytical tools, how to access, etc.
Intellectual property context
Research Scientist
Data organization, Preservation, Management, Personnel needs, etc.

Figure 1. An overview of a Data Curation Profile

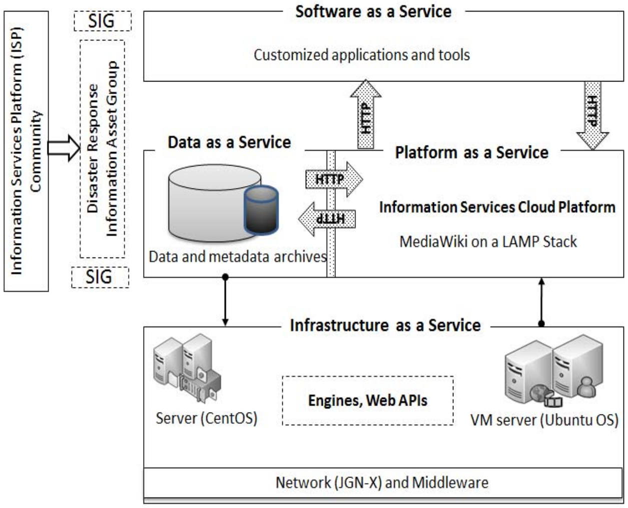

Information Services Platform (ISP) Community
SIG
Disaster Response Information Asset Group
SIG
Software as a Service
Customized applications and tools
HTTP
HTTP
Data as a Service
Data and metadata archives
HTTP
HTTP
Platform as a Service
Information Services Cloud Platform
MediaWiki on a LAMP Stack
Infrastructure as a Service
Server (CentOS)
Engines, Web APIs
VM server (Ubuntu OS)
Network (JGN-X) and Middleware

Figure 2. Architecture of the community-based data curation cloud platform.

However, existing cloud and big data platforms offer little flexibility in terms of supporting collaborative work or data curation activities. In order address this shortcoming and support the disaster response community in their data curation activities, we opted for Mediawiki because of its flexibility, ease of customization, extensibility, and superiority in supporting collaborating editing of web content [22]. Furthermore, wikis are extensively used by the scientific community for collaborative research and annotation of data [23]. However, wikis in general have their own drawbacks (e.g. accessing and linking heterogeneous data sources) which many hinder data curation [22]. Data curation on a wiki goes beyond collaboratively editing of data curation profiles alone. It also involves providing users the tools they would need to enable them search for relevant data, link other data sources to their DCP, upload or register data for their own DCP and for other curators, annotate texts, make data citations, track data provenance [24], and discuss or talk about their research or data curation profile(s). These activities can be mapped to a user data curation profile in order to provide detailed information on particular data forms that are curated by the individual [21].

III. Architecture of the Data Curation Cloud Platform

The architecture of the data curation cloud platform aims to make it easy for the disaster response community (described in more detail in section IV) to curate data without needing to worry about the underlying technical details and the location of the data or services we offer. Furthermore, the community should be able to access our distributed data and services anytime, any-where, so long as the members have access to a computer with a Web browser and Internet connectivity. The disaster response community, shown on top left side of Figure 2, is a special interest group or sub-community of the overall information services platform community. The community uses the infrastructure for open collaboration [25], data curation, and collaborative editing of data curation profiles.

The platform architecture depicted in Figure 2 exhibits typical cloud service models [22]; commonly referred to as infrastructure as a service (IaaS), software as a service (SaaS), and platform as a service (PaaS). In addition, we also offer, on demand, data as a service (DaaS). [26] pointed out that DaaS users can either download or query data from different data assets (through APIs), and they do not need to fetch and store giant data assets and search for the required information in the data asset. In what follows, we summarize how the cloud service models are loosely mapped to our cloud platform infrastructure.

A. Software as a Service - SaaS

This layer consists of customized web-based applications we have developed to support data curation. The applications are accessible over the internet and made available as services on the platform. Applications currently available to the disaster response community are the following:

1) *Search and retrieval systems*. These systems are used by the ISP community to search and retrieve data and metadata for their research. The systems use spatial, temporal, ontological, geographical, and citation correlations between datasets and visually display the search results to help scientists better understand the results of their queries;
2) *Data citation tool*. The data citation tool allows users

to make citations and append the results to their data curation profiles;

3) *Data registry*. The community can use the data registry application to registry or upload data for their research;
4) *Datasets download*. The community can use this tool to package and download whole or parts of information assets for use outside the cloud platform;
5) *Data curation tool (DCT)*. This tool can be used by the community to curate datasets, by combining two or more datasets (e.g. Temperature and rain data) and importing the curation results back to their data curation profile;
6) *Verify Data Provenance (VDP)*. The VDP tool allows data curators to check the provenance information of a particular dataset and assess the risk of linking that dataset with other data;
7) *Quality Assurance (QA)*. Researchers can use the information assets issues tracker to view, vet, and discuss data quality issues;
8) *Collaboration and coordination tools*. Discussion forums and mailing lists are provided to help the community coordinate and collaborate on their big data development activities.

B. Platform/Data as a Service - PaaS/DaaS

This layer consists of the information services cloud platform installed on MediaWiki, data, and metadata archives. These provide the environment for curating data. HTTP request/response gateways gives the community on-demand access to resources, data and metadata, applications, visualizations, and data curation profiles. The community can access data in the form of HTML, XML, audio/video, text, CSV, etc.

C. Infrastructure as a Service - IaaS

The information technology infrastructure of the cloud platform consists of the Japan Gigabit Network[1](JGN-X) node connected to our lab, and physical and virtual machine (VM) servers with their respective operating systems. A scalable and high-performance virtualization VMWare allows us to install and maintain multiple applications on a single physical platform. Physical machines are sandwiched between firewalls and linked directly to the JGN-X network. The infrastructure also contains engines and web APIs that handle visualizations, simulations, and I/O service requests to the data and metadata stored in archives databases.

IV. THE DISASTER RESPONSE RESEARCH COMMUNITY

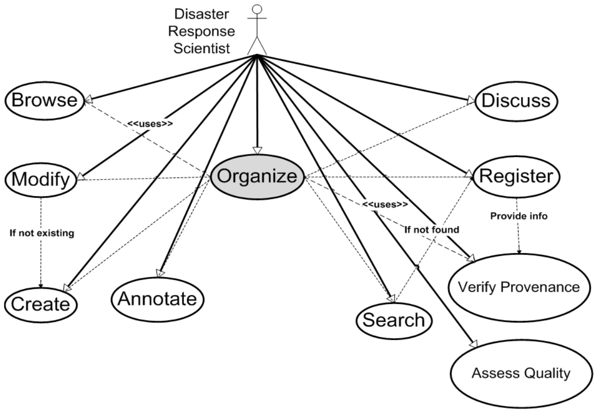


Figure 3. Disaster response scientist Use Cases.

The objective of the information services platform (ISP) lab is to build a cloud platform that will support service consumers and providers in their use and provision of information assets [22]. The community-based cloud platform cater for the needs of a broad range of e-science and data intensive science stakeholders. The community is loosely sub-divided into subcommunities or special interest groups that collaborate and work on topics related to Cyber-Physical systems, knowledge language grid applications, web archives and data citations, and disaster response information assets [22]. This research focuses on the latter group, with emphasis on disaster response scientists.

The function of the disaster response information asset group is to individually and collectively create information assets for analyzing situations and responses relating to target disasters (e.g. typhoons, hurricanes, earthquakes, etc.). The user profile of the group consists of disaster response scientists, humanitarian relief organizations and volunteers, and the general public who are just interested in obtaining and understanding (shallow) disaster situations.

A. Use Case Scenarios for Big Data Development

Figure 3 depicts ten use cases for one member of the disaster response information assets group - the disaster response scientist. Dotted lines indicate cause-effect relationship between the use cases. A detail description of each use case and how it is implemented on the cloud platform is shown in Table I. As shown at the center of Figure 3, the *organize* use case is central representation of how we expect a disaster response scientist to interact with the cloud platform, as well as the information assets on it. On the platform, a scientist can *browse* recently changed data curation profiles, profiles that were cited most by other researchers, and most cited datasets. This use case will help users understand disaster situations being discussed by the community. The scientist can login and edit or *modify*, *annotate* or comment on any of the data curation profile. Any changes made (additions/deletions) can be tracked, viewed, and/or reverted through revision histories.

Furthermore, a scientist can *search* for particular disaster topics (e.g. earthquakes, floods, tsunamis), keywords (e.g. heavy rain), or datasets related to specific disasters, in a

[1] http://www.jgn.nict.go.jp/english/index.html

Table I
DISASTER RESPONSE SCIENTISTS USE CASES AND THEIR IMPLEMENTATION.

Use Case	Implementation on the cloud platform (Ref, Figure 2)
Create	Create new IA in the form of a DCP, upload or register a new datasets, or write a wiki page on a specific disaster.
Search	Search for wiki pages, DCPs, and datasets that are relevant to a specific disaster (e.g. Earthquake) using customized search systems.
Organize	Organize the retrieved IAs according correlations between the disaster (e.g. Earthquake) and other influences or factors (e.g. infrastructure damage, loss of life, etc.).
Browse	Browse the organized IAs to understand a particular disaster situation.
Modify	Edit or modify organized IAs to improve their quality and usability.
Annotate	Annotate or comment on the quality and usability of IAs.
Provenance	Verify the provenance [24], [27] of IAs.
Quality	Assess the quality of big data information assets by using quality assurance issue tracker.
Register	Register or upload data to the cloud platform, save edited data curation profiles, wiki pages for future use.
Discuss	Discuss or talk about disaster response in a particular DCP page.

- IA = Information Asset, DCP = Data Curation Profile

given region (e.g. satellite maps of Tohoku (Japan) earthquake). If the disaster topic or dataset, for instance, is not found in the *organized* or existing information assets, the scientist can *create* his own topic to address contemporary or past disaster situations. If the dataset is not available in the bid data repository, the scientist can *register* or upload datasets related to the disaster he is interested in. In registering or uploading datasets, users will be required to provide additional information (e.g. Name of asset and uploader, date, license, brief summary describing the asset, etc.) that will later be used to verify the *provenance* of the information asset. Using the information assets quality assurance issue tracker, users can assess the *quality* of the information assets on the platform. All platform users have access to *discussion* forums and mailing list that they can use to reach-out to other users, discuss about quality, modifications, and usage of information assets on the platform.

V. CURATING BIG DATA INFORMATION ASSETS

The data curation model in Figure 4 demonstrates how the disaster response community is involved in the process of curating data on the cloud platform. The model contains three key components:

- *Actors*: Actors in the data curation process are
  1) Data providers; a disaster response scientist who can register or upload data to the cloud platform.
  2) Data curators; disaster response scientist who can create a data curation profile, curate or combine two or more datasets, and organize the information in such a way that others can understand a particular disaster situation. For example, a scientist can combine, correlate, and create visualization of typhoon and twitter data to help other platform users have a gist of what people in the typhoon affected region are saying (e.g. where to find shelter, medical help, evacuation spots, etc.).
  3) Data consumers; disaster response scientists who just browse (use case discussed above) or lurk on the cloud platform to understand a particular disaster situation.

     In this group of actors, we also have *prosumers* (data providers who are also consumers of their own data), *curosumers* (data curators who are also consumers of their own curated data), and *wikilators* (people who facilitate data curation activities on the wiki platform).

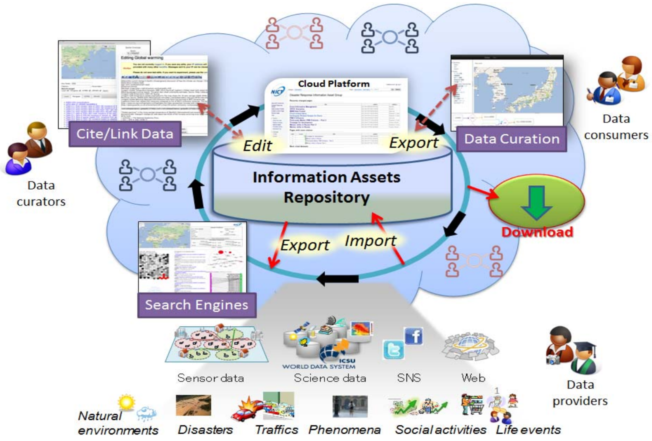
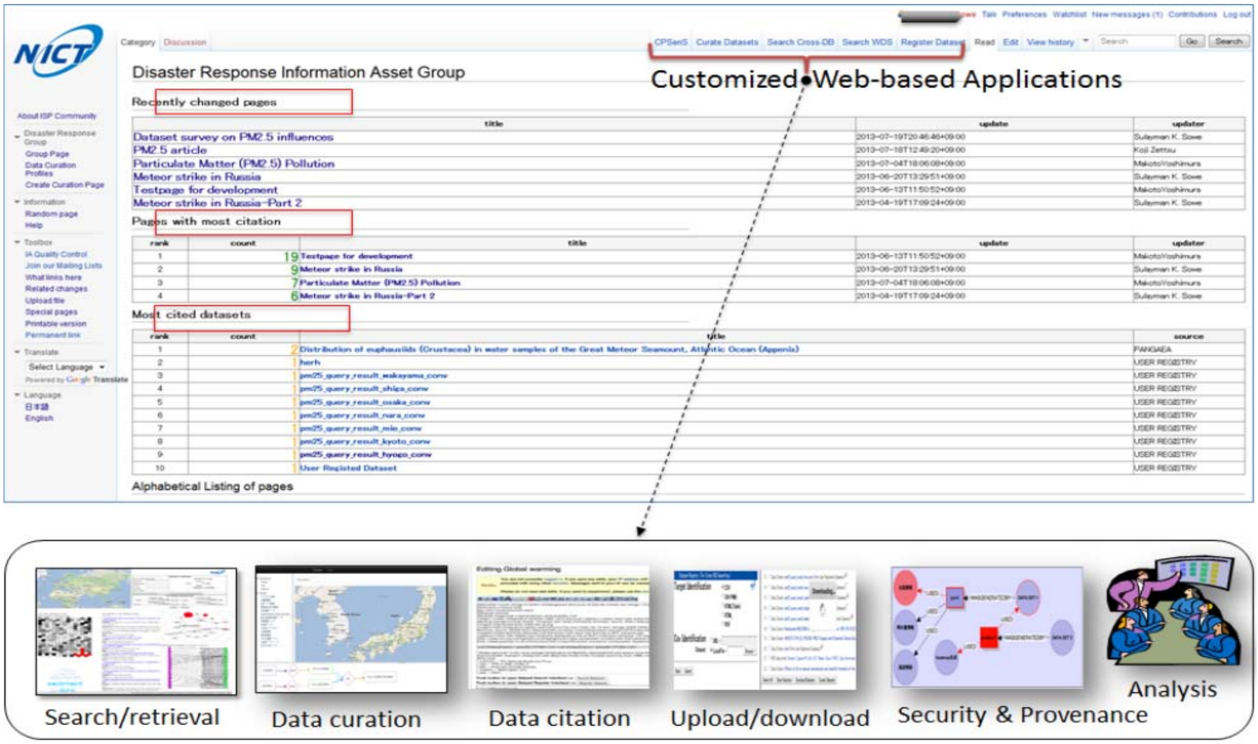


Figure 4. Data curation model.

- *Data*: As shown at the bottom of Figure 4, the cloud platform initially contains data assets; such as sensor data, science data, social sensing data, web data archives, natural environment data (e.g. rain, temperature), disaster and traffic data, etc. As of January 2013, the ISP lab archived 2.4PB of data, and this is said to increase every three months or so.
- *Platform and Tools*: The cloud platform contains customized tools that allow the actors to search for disaster related datasets, cite or make references by linking their work or profiles with that of others and with other datasets, and tools for registering and curating datasets.

In the cyclic nature of curating data on the platform (center of Figure 4), data is stored in the information assets repository in the cloud and can be accessed by actors on demand. Actors *export* or check-out datasets from the repository using customized tools. They can cite the dataset, embed it in wiki pages they are editing, or curate the dataset using the data curation tool. Actors can do this individually on their workspace or in collaboration with other researchers on the platform. At this stage, the dataset that was originally



Disaster Response Information Asset Group

Recently changed pages

title	update	updater
Dataset survey on PM2.5 influences	2013-07-19T20:46:46+09:00	Sulayman K. Sowe
PM2.5 article	2013-07-18T12:49:20+09:00	Koji Zettsu
Particulate Matter (PM2.5) Pollution	2013-07-04T18:06:08+09:00	MakotoYoshimura
Meteor strike in Russia	2013-06-20T13:29:51+09:00	Sulayman K. Sowe
Testpage for development	2013-06-13T11:50:52+09:00	MakotoYoshimura
Meteor strike in Russia-Part 2	2013-04-19T17:09:24+09:00	Sulayman K. Sowe

Pages with most citation

rank	count	title	update	updater
1	19	Testpage for development	2013-06-13T11:50:52+09:00	MakotoYoshimura
2	9	Meteor strike in Russia	2013-06-20T13:29:51+09:00	Sulayman K. Sowe
3	7	Particulate Matter (PM2.5) Pollution	2013-07-04T18:06:08+09:00	MakotoYoshimura
4	6	Meteor strike in Russia-Part 2	2013-04-19T17:09:24+09:00	Sulayman K. Sowe

Most cited datasets

rank	count	title	source
1	2	Distribution of euphausiids (Crustacea) in water samples of the Great Meteor Seamount, Atlantic Ocean (Appenix)	PANGAEA
2	1	herh	USER REGISTRY
3	1	pm25_query_result_wakayama_conv	USER REGISTRY
4	1	pm25_query_result_shiga_conv	USER REGISTRY
5	1	pm25_query_result_osaka_conv	USER REGISTRY
6	1	pm25_query_result_nara_conv	USER REGISTRY
7	1	pm25_query_result_mie_conv	USER REGISTRY
8	1	pm25_query_result_kyoto_conv	USER REGISTRY
9	1	pm25_query_result_hyogo_conv	USER REGISTRY
10	1	User Registed Dataset	USER REGISTRY

Alphabetical Listing of pages

Figure 5. Web portal of the disaster response community, showing custom data curation applications.

exported form the repository would have gone through "many hands" and is being improved upon along the data curation cycle in such a way that, we can say, it has more value and is of "better quality" than the original dataset that was exported from the repository.

Researchers can also select, package, and download whole or part of the curated data. Alternatively, they can, altruistically, *import* or check-in the new and improved information asset to the information assets repository. This cycle continues in perpetuity, creating more information assets for the community. Thus, after many iterations, we posit that the information assets repository with its 2.4PB of data will, indeed, contain more community generated (big) data.

VI. Implementation and Discussion

As discussed in the SaaS layer of the architecture of the cloud platform, figure 5 is a screenshot of the disaster response community web portal showing the customizations we added to the MediaWiki core. With these customization tabs the group can discuss their data curation profiles or research activities in the designated wiki pages, search the wiki using both the search systems and insert the search results into the wiki pages, register or upload/download datasets to and from the platform. Furthermore, on the front page of the website, visitors can see recently changed or curated profiles pages, rank of data curation profile pages with most citations, and a rank of the most cited datasets in each data curation profile. The menu items on the left shows, among others, users can create their own data curation profile using custom templates, assess the quality of the information assets by clicking the "IA Quality Control" link, and join the community mailing lists, and data curation documentation and FAQs.

Currently the disaster response community consists of 19 members (research scientists and engineers). Together, they have created 138 wiki pages in two weeks. This community is involved in the testing and experimentation phase of the platform. They community generated the initial wiki contents that could then be used to solicit further contributions from a larger community of users. They revised and discussed the prototyped contents, and collaboratively edited the sample data curation profiles. The requests for improvements and enhancement filed in by the community provided a feedback loop in our community development process.

A. Generalizability and transferability of findings

Data curation practices are highly specific to individual communities and, therefore, may vary widely across communities such as disaster response, library and information sciences, and even in condense matter physics [28]. However, while the context of this research is unique to our research environment, we conjecture that the disaster response use cases and the cloud architecture, with a strong emphasis on the user community, can serve as a starting point for eliciting functional and non-functional requirements for cyber technologies that provide ubiquitous services for users in general, and scientific communities in particular. Furthermore, the data curation model can be easily transferred in another context or discipline to model data curation activities in other domains.

B. Limitations

As far as supporting communities curate data is concern, this research has certain limitations that must be mentioned. Some of these issues include: discussion on the maintainability and scalability of the platform, how we guarantee the security [29] and intellectual property of the big data information assets, how the provenance of the registered and curated data are verified, unified format for our registered datasets, the metrics for assessing the quality of big data information assets, licensing of the curated data [30], and how to motivate the disaster response scientists taking part in the data curation process. These are important issues that we hope will generate more discussions amongst conference participants.

VII. Conclusion

In this paper, we presented and discussed the architecture of a cloud platform that supports disaster response scientists curate big data information assets. The architecture combines the flexibility of cloud computing with extensive customization of MediaWiki. While the architecture looks like any other cloud computing infrastructure, we placed great emphasis on the community layer, and discussed how the disaster response communities can leverage the SaaS, DaaS, and PaaS layers for their data curation activities. We presented use cases for disaster response scientists and described how these are implemented on the cloud platform. A data curation model was introduced to demonstrate how the collaborative development of data curation profiles could lead to the generation of more (big) data for the community. In our implementation and discussion section, we showed how the disaster response web portal was designed to meet the needs of data curation, and what the disaster response community can do on the website.

Supporting scientific communities in their data curation activities is a complicated process that requires an interdisciplinary effort. Despite some of the research limitations we mentioned, working with the small community of nineteen testers shows that, as a proof of concept, it is feasible to design, develop, and maintain a platform that can support disaster response scientists curate big data information assets.

Agile software development is the hallmark of our effort to provide a sustainable cloud platform to support data curation in various domains. We consider this research to be the beginning of a dialogue that will lead to better understanding of big data community dynamics, big data community metrics, and how to develop cyber technologies that can support research communities. Research in the fields of cloud computing, cloud resource management and allocation, QoS, big data infrastructure development, scientific data management, cyber-physical cloud computing, data curation and digital libraries, Open Source software development, all have a role to play in this dialogue. And the 5th International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC 2013) provides a fertile ground to begin this dialogue.

References

[1] M. J. Bietz, A. Wiggins, M. Handel, and C. Aragon, "Data-intensive collaboration in science and engineering," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*, 2012, pp. 3–4.

[2] R. Heimann, "Big social data. the long tail of science data," *Imaging Notes*, vol. 28, no. 1, pp. 34–37, 2013.

[3] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digi tal shadows, and biggest growth in the far east," International Data Corporation (IDC), Tech. Rep., December 2012.

[4] A. Cuzzocrea, I.-Y. Song, and K. C. Davis, "Analytics over large-scale multidimensional data: the big data revolution!" in *Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP*, 2011, pp. 101–104.

[5] boyd danah and K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," *Information, Communication, & Society*, vol. 15, no. 5, pp. 662–679, 2012.

[6] S. K. Sowe, L. Angelis, I. Stamelos, and Y. Manolopoulos, "Using Repository of Repositories (RoRs) to Study the Growth of F/OSS Projects: A Meta-analysis Research Approach," in *In Third International Conference on Open Source Systems*. Springer, 2007, pp. 11–14.

[7] J. Lin and D. Ryaboy, "Scaling big data mining infrastructure: the twitter experience," *SIGKDD Explor. Newsl.*, vol. 14, no. 2, pp. 6–19, Apr. 2013.

[8] M. W. A. F. Ian Mitchell, Mark Locke, *The White Book of Big Data.* Fujitsu Services Ltd., 2012.

[9] I. Shklovski, L. Palen, and J. Sutton, "Finding community through information and communication technology in disaster response," in *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, 2008, pp. 127–136.

[10] P. Hunter, "Journey to the centre of big data," *Engineering Technology*, vol. 8, no. 3, pp. 56–59, 2013.

[11] A. Erdman, D. Keefe, and R. Schiestl, "Grand challenge: Applying regulatory science and big data to improve medical device innovation," *Biomedical En-*

gineering, IEEE Transactions on, vol. 60, no. 3, pp. 700–706, 2013.

[12] V. Marx, “Biology: The big challenges of big data,” *Nature*, vol. 498, p. 255?260, June 2013.

[13] N. Elmqvist and P. Irani, “Ubiquitous analytics: Interacting with big data anywhere, anytime,” *Computer*, vol. 46, no. 4, pp. 86–89, 2013.

[14] J. W. Lee, J. Zhang, A. S. Zimmerman, and A. Lucia, “DataNet: An emerging cyberinfrastructure for sharing, reusing and preserving digital data for scientific discovery and learning,” *AIChE Journal*, vol. 55, no. 11, pp. 2757–2764, 2009.

[15] M. Rost, L. Barkhuus, H. Cramer, and B. Brown, “Representation and communication: challenges in interpreting large social media datasets,” in *Proceedings of the 2013 conference on Computer supported cooperative work*, 2013, pp. 357–362.

[16] S. K. Sowe, I. Stamelos, and L. Angelis, “Understanding Knowledge Sharing Activities in Free/Open Source Software Projects: An Empirical Study,” *Journal of Systems and Software*, vol. 81, no. 3, pp. 431–446, March 2008.

[17] DCC, “What is digital curation? Digital Curation Centre,” http://www.dcc.ac.uk/digital-curation/what-digital-curation, june 2013.

[18] S. Choudury, “Data curation: An ecological perspective,” *College & Research Libraries News*, vol. 71, no. 4, pp. 194–196, April 2010.

[19] P. Lord, A. Macdonald, L. Lyon, and D. Giaretta, “From data deluge to data curation.” JISC (the Joint Information Systems Committee) and the UKfs e-Science Core Programme, Tech. Rep., 2003. [Online]. Available: http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/150.pdf

[20] L. Y. Karen S. Baker, “Data stewardship: Environmental data curation and a web-of-repositories,” *International Journal of Digital Curation*, vol. 4, no. 2, pp. 12–27, 2009.

[21] M. Witt, J. Carlson, S. Brandt, and M. Cragin, “Constructing data curation profiles,” *International Journal of Digital Curation*, vol. 4, no. 3, pp. 93–103, 2009.

[22] S. K. Sowe, K. Zettsu, and Y. Murakami, “A Model for Creating and Sustaining Information Services Platform Communities: Lessons learnt from Open Source Software.” in *The 5th ITU Kaleidoscope conference on Building Sustainable Communities, Kyoto, Japan*, 2013, pp. 13–20.

[23] M. Waldrop, “Big data: Wikiomics,” *Nature*, vol. 455, pp. 22–25, 3 September 2008.

[24] S. Davies, “Still building the memex,” *Magazine, Communications of the ACM*, vol. 54, no. 2, pp. 80–88, 2011.

[25] D. Riehle, J. Ellenberger, T. Menahem, B. Mikhailovski, Y. Natchetoi, B. Naveh, and T. Odenwald, “Open collaboration within corporations using software forges,” *Software, IEEE*, vol. 26, no. 2, pp. 52–58, 2009.

[26] Q. Hieu, T. Pham, and H. Truong, “Demods: A description model for data-as-a-service,” in *Advanced Information Networking and Applications (AINA), 2012 IEEE 26th International Conference on*, March 2012, pp. 605–612.

[27] L. Di, Y. Shao, and L. Kang, “Implementation of geospatial data provenance in a web service workflow environment with iso 19115 and iso 19115-2 lineage model,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2013.

[28] C. Hinnant, B. Stvilia, S. Wu, A. Worrall, K. Burnett, G. Burnett, M. Kazmer, and P. Marty, “Data curation in scientific teams: an exploratory study of condensed matter physics at a national science lab,” in *Proceedings of the 2012 iConference*, 2012, pp. 498–500.

[29] P. K. Manadhata, “Big data for security: challenges, opportunities, and examples,” in *Proceedings of the 2012 ACM Workshop on Building analysis datasets and gathering experience returns for security*, 2012, pp. 3–4.

[30] A. Ball. (2012) How to License Research Data. DCC How-to Guides. Edinburgh: Digital Curation Centre. [Online]. Available: http://www.dcc.ac.uk/resources/how-guides