PROCEEDINGS OF SCIENCE

Capturing User-Generated Metadata

Wataru Takase[*]
KEK
E-mail: wataru.takase@kek.jp

Adil Hasan
University of Liverpool
*E-mail:*adilhasan2@gmail.com

Yoshimi Matsumoto
KEK
E-mail: yoshimi.matsumoto@kek.jp

Takashi Sasaki
KEK
E-mail: takashi.sasaki@kek.jp

A key component of the management of digital data is the collection of metadata. Metadata is defined as data about data. It encompasses the description of the data and is essential in using or reusing the data. Metadata is often collected at the time the data are produced, but little attention is paid to the metadata that is created through the use or attempted use of the data. In this paper we propose an approach to enhance data management through the collection and management of metadata produced during the use or reuse of the data. The paper also describes an example of the approach applied to the iRODS data management system.

The International Symposium on Grids and Clouds (ISGC) 2014,
March 23-28, 2014
Academia Sinica, Taipei, Taiwan

[*] Speaker.

1. Introduction and Related Work

An important part of the creation of digital data is the creation of information that describes the data. This information is called metadata [1] and is critical to the successful discovery and use of the data. There exist many different types of metadata and all these can be grouped into three broad categories [2]: Descriptive metadata which contains all the information necessary to discover and understand the data. Metadata that fit within this category could be the title or abstract of the dataset. Provenance information which describes the pedigree of the data can also be classified as descriptive information that covers a description of how the dataset was created which helps in understanding how the dataset can be used. The second category is Structural metadata which covers how the data are arranged within the file. For example the column titles in a tabular dataset. Structural metadata is necessary to correctly use the data. The final category is Administrative metadata that is necessary for managing the data which includes technical information to ensure the integrity of the data as well as information on who can access the data.

In this paper we focus mainly on the descriptive metadata. The categorization of descriptive metadata has been extensively studied and has resulted in standards such as the Dublin Core Metadata Element set (DCME) [3] which describes the most basic, common metadata elements needed to describe the data. Many domain-specific de-facto standards have been developed that derive from DCME (see for example [4]). These standards have been used to develop national, domain-specific and local repositories that enable the collection of descriptive metadata (see for example [5], [6], [7]) at the time of data deposit. Some of these repositories employ automated tools that have been developed to harvest metadata from other sources (usually other repositories) using the standard Open Archive Initiative - Protocol for Metadata Harvesting (OAI-PMH [8]).

Work on folksonomies [9] addresses user-defined metadata through the mechanism of tagging data. The tags can cover relevant features in the dataset. This enables data discovery by allowing terms meaningful to the user community to be associated to the data. Lu et al. [10] have compared the effectiveness of metadata provided by the author of the data with that provided by the users of the data. They observed user-generated metadata to be much more effective for enhancing web clustering performance. Our approach also starts from the premis that user-defined metadata in addition to that provided by the creators of the data is essential to the effective reuse of the data. We describe a practical approach to collect user-generated metadata which transforms the metadata collection process from a push-model where the potential use of the data is anticipated by the data creator to a pull-model where the use of the data drives the collection of metadata.

In Section 2 we briefly describe the metadata lifecycle and Section 3 describes our approach to demand-driven metadata collection. In Section 4 we describe an example of the approach using the iRODS data management system. A summary and future work is described in Section 5.

2. The Metadata lifecycle

Taking inspiration from the data lifecycle [11] we can define a lifecycle for metadata as shown in Figure 1. The cycle starts with the *identify metadata* step. This step is similar to the data lifecycle creation phase and consists of the data creators and user community defining the important information to enable data use. The step may make use of metadata standards that the community

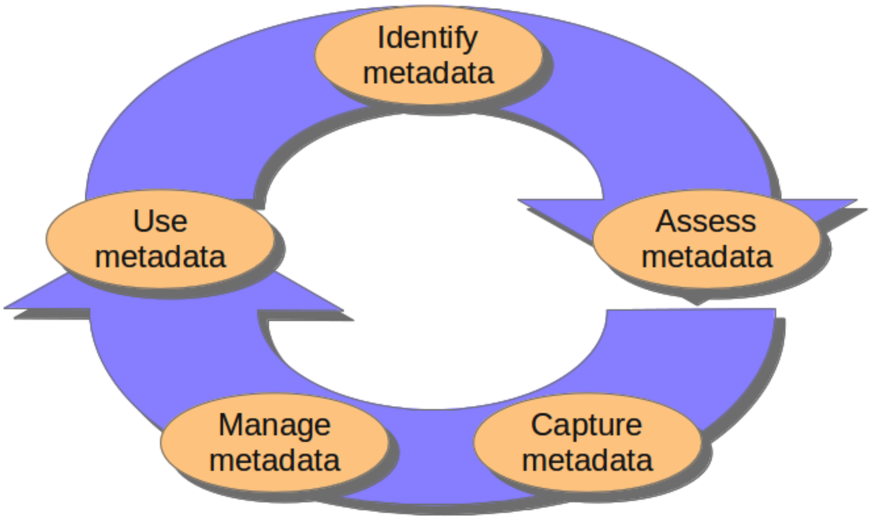


Figure 1: The metadata lifecycle.

and the repository accept. The *assess metadata* step is similar to the appraisal phase of the data lifecycle. It entails matching the needs of the community with those of the repository. It is also guided by the relevant metadata standards. The *capture metadata* step may entail manual annotation of the data or automated metadata harvesting. This step is equivalent to the ingest phase of the data lifecycle. The *manage metadata* step covers the preservation action and storage phases of the data lifecycle. The metadata is stored in structured storage systems (such as a relational database). Management also entails ensuring the integrity and accessibility of the metadata. It may also require migration of the metadata from one storage system to another. The final step, *use metadata*, addresses the access and use of the metadata either by persons or services. This is the same as the access, use and reuse phase of the data lifecycle. It covers the access by authorized users as well as repackaging the metadata (for example in an exchange format) for use or reuse.

Ya-Ning, Chen and Lin [12] also worked on a definition of a lifecycle model and methodology for metadata. However, their approach primarily focuses on the acquisition of metadata management systems and less on the lifecycle of the metadata itself.

The initial identification and collection of metadata is well established and is an important part of many systems (see for example the Open Archival Information System [2]). There is a recognition within the community of the importance of accommodating regular updates of metadata that is well understood and has been addressed by the community (see for example [13]). However, there is an important metadata component that has largely not been covered to date: the metadata generated by the process of understanding the data. We describe this metadata and the approach in the following section.

3. Demand-Driven Metadata

New metadata can be generated as interested parties make use of the data. For example, consider a collection of raw data (i.e. the direct recordings of the experiment) from a CERN LHC experiment. A further study of these raw data by researchers may result in the observation of new phenomena (such as the observation of evidence for the Higgs boson). Metadata on these phenomena (for example the event number, the type of phenomena etc) could then be added to the existing set of raw data and enable future researchers the opportunity to study in further detail those phenomena.

Current approaches to metadata management take into account the collection of this type of author-defined metadata. There is another type of metadata that arises from the process of understanding the data that is rarely collected. This type of metadata can best be understood by means of an example.

Consider a paper published by one group of researchers reporting the observation of the Higgs Boson. We can view the paper as a collection of data which are the parameters of the boson (e.g. the mass and decay modes etc) and metadata which is the surrounding description of how the parameters were obtained. The creators of the paper have in mind potential users of the paper such as fellow experimentalists, phenomenologists and theorists and provide description that they anticipate will be sufficient for these users. Once the paper has been published keen researchers study the paper in order to make use of the data.

In most cases the contents of the paper are sufficient for researchers to reuse the data. But, in some cases the details contained in the paper are insufficient. For example, fellow experimentalists who intend to combine the parameters with those from other experiments may require more details on the definition of the errors than are in the paper in order to combine the results correctly. In these cases the researcher will contact the authors of the paper in order to clarify ambiguous points or to obtain missing information. We argue that the dialogue is also an important piece of metadata that should be captured and kept with the data. The dialogue that is captured can also cover discussions of features of a dataset. These discussions essentially form annotations of the dataset. Our belief is backed-up by empirical evidence: the popularity of sites such as StackExchange [14] which contains questions and answers on a variety of programming and software issues. These questions and responses can be viewed as supplementary metadata for the use of programming languages and software libraries. The Q&A forum has become so popular that it now includes a wide-variety of activities such as: mathematics, bicycles, languages and travel. In the following section we describe a demonstration of our approach.

4. Demand-Driven Metadata Demonstrator

In this section we describe the demonstrator we have created to illustrate the capture of demand-driven metadata. We have developed an application *Gyoza* that interfaces to the integrated Rule Oriented Data System (iRODS [15]). We describe the architecture and describe how it can be applied.

4.1 iRODS

iRODS is a data grid software system and provides a logical file system connecting geographically distributed physical devices [15]. A mapping between the logical namespace and physical location is maintained in a database that serves as the iRODS metadata catalog named iCAT. Each object stored in iRODS has system-defined metadata such as data owner name, physical data path and data checksum value. An iRODS system consists of an iCAT, an iRODS server that interfaces to the iCAT and zero or more iRODS storage servers. The owner of the data can also add owner-defined metadata to each object which is registered in iRODS.

The iRODS Rule Engine allows policies to be enforced on the data [15]. Rules can be written for many different types of policies. Rules can be invoked automatically in response to certain conditions or triggers and can also be invoked on the commandline. Many researchers have investigated the use of iRODS to aid in the long-term management of digital data through the management of metadata as well as migration from one data management system to another [6, 16, 17, 18].

4.2 Gyoza

We have developed a web interface (called Gyoza) to capture user-generated metadata for data stored in iRODS. Gyoza wraps iRODS and provides a discussion interface. iRODS users can post messages about iRODS data and metadata to Gyoza and data owners or experts can post responses. Gyoza captures such dialogue between the owner/expert of the data and user of the data. iRODS user can refer to iRODS metadata and user-generated metadata through the Gyoza interface. Goza provides the following functions: add missing information about data and metadata, ask data owner about metadata, and allow people to refer to past discussions.

Figure 2 shows the Gyoza architecture. It consists of four components: a user interface which handles user's requests and displays web pages, a translator which translates posted message to metadata schema used by the repository, storage which serves as the user-generated metadata store, and a connector which communicates with iRODS using PRODS [19]. Gyoza stores the message dialogue as a thread which is linked to iRODS data. Each posted messages is associated to thread.

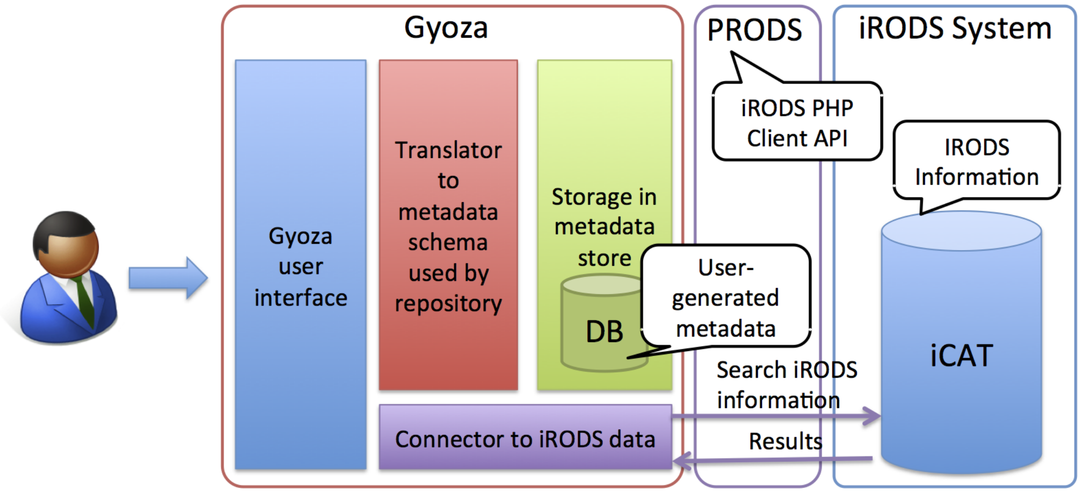


Figure 2: Gyoza architecture.

Figure 3 shows the Gyoza usage workflow. At first the user needs to login to Gyoza using an iRODS account. Once logged in the user can search existing threads and iRODS data by keywords and can also check threads they are a member of. When a user initiates a request, Gyoza fetches information on the thread from its own database. Gyoza also fetches iRODS information from iCAT and then returns the result on the web. The user can view the returned thread, add new comments to it or create a new thread if they have new questions about the data or metadata.

Figure 4 shows the detailed information returned by a query. The page displays three types of metadata: iRODS system metadata, owner-defined metadata, and threads of user-generated metadata.

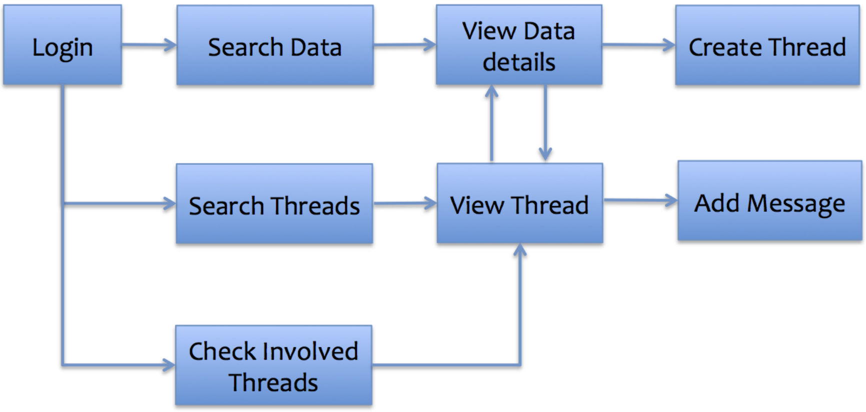
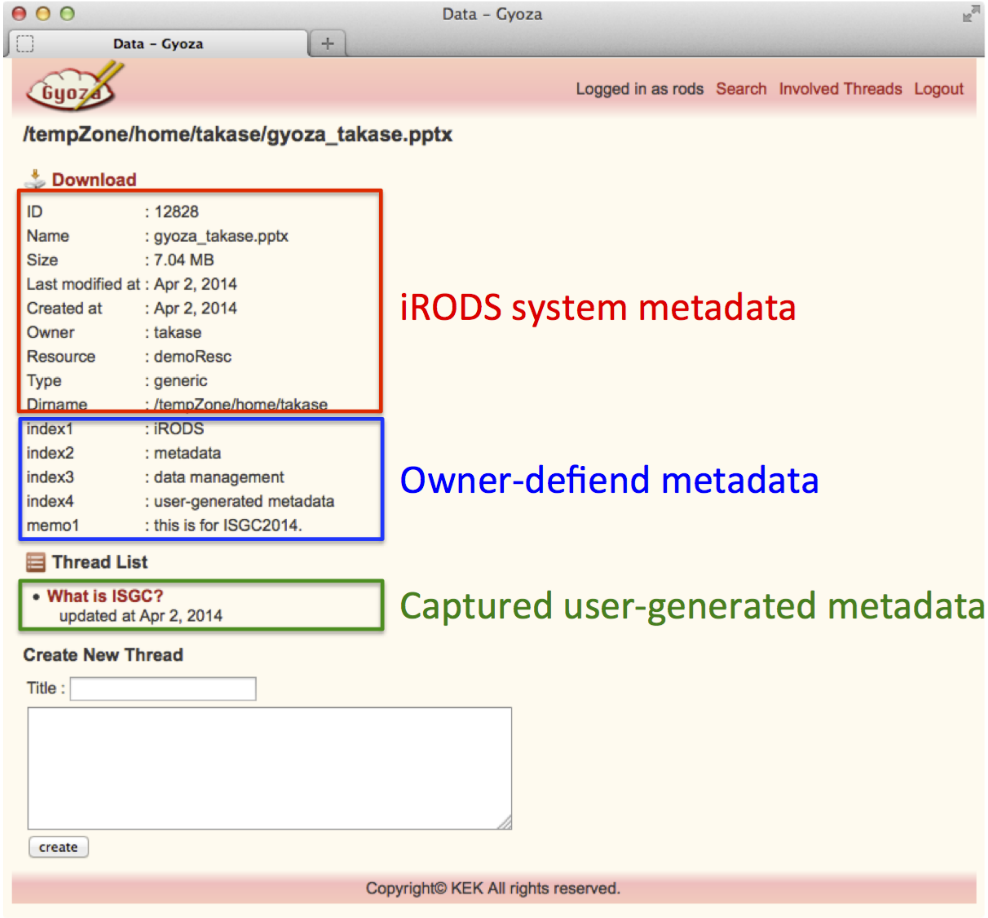


Figure 3: Gyoza usage workflow.



/tempZone/home/takase/gyoza_takase.pptx

ID : 12828
Name : gyoza_takase.pptx
Size : 7.04 MB
Last modified at : Apr 2, 2014
Created at : Apr 2, 2014
Owner : takase
Resource : demoResc
Type : generic
Dirname : /tempZone/home/takase

iRODS system metadata

index1 : iRODS
index2 : metadata
index3 : data management
index4 : user-generated metadata
memo1 : this is for ISGC2014.

Owner-defiend metadata

- What is ISGC?
  updated at Apr 2, 2014

Captured user-generated metadata

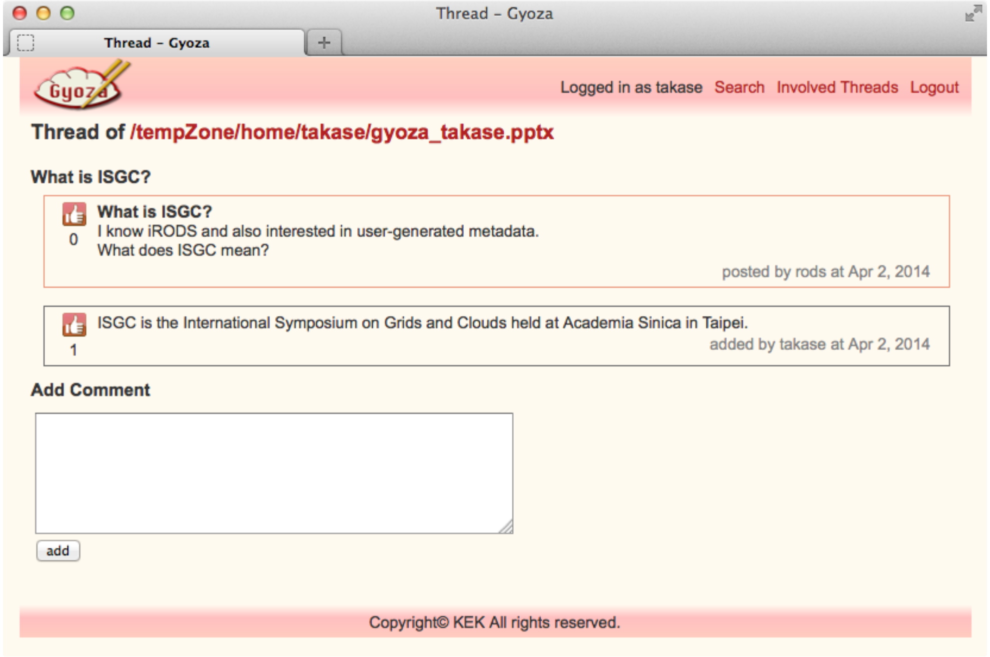
Create New Thread

Title :

Copyright© KEK All rights reserved.

Figure 4: Gyoza: Data details page.

In Figure 4 we show an screenshot of a user creating a new thread. An indication of the quality and relevance of each response can be determined by the number of up-votes. When a user creates a new thread or posts a message, Gyoza translates the user's request into the Gyoza metadata scheme and then stores it to own database. The Gyoza metadata scheme is a subset of the Dublin Core and contains the creator, title, message, creation date.



Thread of /tempZone/home/takase/gyoza_takase.pptx

What is ISGC?

0 What is ISGC?
I know iRODS and also interested in user-generated metadata.
What does ISGC mean?
posted by rods at Apr 2, 2014

1 ISGC is the International Symposium on Grids and Clouds held at Academia Sinica in Taipei.
added by takase at Apr 2, 2014

Add Comment

Copyright© KEK All rights reserved.

Figure 5: Gyoza: Viewing thread page.

5. Future Work

The approach we have described covers the capture of the dialogue between the creator of the dataset and the user of the dataset with a user-driven method of scoring to weight the information. Such an approach, if unchecked, is prone to misuse. Our approach has the implicit requirement that a *data custodian* or *data manager* that has some detailed knowledge of the dataset helps to maintain the dataset. This means that the weighting and comments would need to be reviewed by the data manager regularly to ensure they are relevant and reflect the uses of the dataset. This requires the development of an administrator interface that allows the data manager to approve or reject dialogue metadata as meaningful or misleading respectively. Keeping a history of the accepted and rejected metadata would also be of potential value to later users of the dataset. And it is possible to implement a feature-based indexing (see for example [20]) on the captured dialogue that would provide a more feature-rich means of searching the metadata.

6. Summary

In this paper we have proposed a demand-driven metadata approach to enhance data management. People currently collect metadata when data are created. And, they collect metadata when data are used. However, the metadata that comes from the process of understanding the data is usually not collected. That is important metadata. So, we propose to capture this metadata and associate it to the data so when other users come along they can look at these chats which may help them to understand the data.

Our approach has been demonstrated by Gyoza. Gyoza wraps existing iRODS system and captures user-generated metadata. Gyoza is useful for adding missing information, taking memo, asking owner, browsing discussion logs.

For future work we intend to evaluate our approach with some existing services. We also need to address the challenge of long-term management of the metadata (which in our opinion requires a domain expert to be associated to the data during its lifetime). We also plan to investigate approaches to ensure the management of the metadata is as low as possible.

Acknowledgments

We thank the generous help of the iRODS developers: Mike Wan and Wayne Schroeder. A. H. is grateful to the support of the KEK foreign researcher grant that enabled this work to be done.

References

[1] NISO, *Understanding Metadata*, Tech. Rep. ISBN: 1-880124-62-9, National Information Standards Organization, 2004.

[2] *Reference Model for an Open Archival Information System (OAIS) Magenta Book*, 2012. `http://public.ccsds.org/publications/archive/650x0m2.pdf`.

[3] "DCMI Home: Dublin Core Metadata Initiative (DCMI)." `http://dublincore.org/`.

[4] "Disciplinary Metadata | Digital Curation Centre." `http://www.dcc.ac.uk/resources/metadata-standards`.

[5] J. Frew and R. Bose, *Earth System Science Workbench: A Data Management Infrastructure for Earth Science Products*, in *Proceedings of the 13th International Conference on Scientific and Statistical Database Management*, pp. 180–189, IEEE, 2001.

[6] D. Walling and M. Esteva, *Automating the Extraction of Metadata from Archaeological Data Using iRods Rules*, *IJDC* **6** (2011), no. 2 253–264.

[7] "Digital Repository Federation (DRF)." `http://drf.lib.hokudai.ac.jp/drf/index.php?Digital Repository Federation (in English)`.

[8] C. Lagoze and H. V. de Sompel, *The Open Archives Initiative: Building a low-barrier interoperability framework*, in *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, pp. 54–62, ACM, 2001.

[9] G. Smith, *Folksonomy: social classification*, 2004. `http://atomiq.org/archives/2004/08/folksonomy_social_classification.html`.

[10] C. Lu, J. ran Park, X. Hu, and I.-Y. Song, *Metadata Effectiveness: A Comparison between User-Created Social Tags and Author-Provided Metadata*, in *HICSS*, pp. 1–10, IEEE Computer Society, 2010.

[11] "DCC Curation Lifecycle Model."
`http://www.dcc.ac.uk/resources/curation-lifecycle-model`.

[12] C. Ya-Ning, C. Shu-Jiun, and S. C. Lin, *A metadata lifecycle model for digital libraries: methodology and application for an evidence-based approach to library research*, in *IFLA Council and General Conference*, 2003.

[13] T. R. Bruce and D. I. Hillmann, *The Continuum of Metadata Quality: Defining, Expressing, Exploiting*, 2004. `http://www.ecommons.cornell.edu/handle/1813/7895`.

[14] "Stack Exchange - Free, Community-Powered Q&A." `http://stackexchange.com/`.

[15] "iRODS." `http://irods.org/`.

[16] J. H. Ward, A. de Torcy, M. Chua, and J. Crabtree, *Extracting and Ingesting DDI Metadata and Digital Objects from a Data Archive into the iRODS Extension of the NARA TPAP Using the OAI-PMH*, in *eScience*, pp. 185–192, IEEE Computer Society, 2009.

[17] J. H. Ward, H. Xu, M. C. Conway, T. G. Russell, and A. de Torcy, *Using Metadata to Facilitate Understanding and Certification of Assertions about the Preservation Properties of a Preservation System*, in *MTSR* (E. Garoufallou and J. Greenberg, eds.), vol. 390 of *Communications in Computer and Information Science*, pp. 87–98, Springer, 2013.

[18] R. Moore and A. Rajasekar, *Evolving Persistent Archives and Digital Library Systems: Integrating iRods, Cheshire3 and Multivalent*, *IJDC* **8** (2013), no. 2 47–67.

[19] "PRODS (iRODS PHP Client API) Dcumentation."
`https://wiki.irods.org/prods_doc/`.

[20] N. Zhang, M. T. Özsu, I. F. Ilyas, and A. Aboulnaga, *Fix: Feature-based indexing technique for xml documents*, in *Proceedings of the 32nd VLDB Conference*, pp. 259–270, 2006.