Parallel Identities for Managing Open Government Data

James P. McCusker, Timothy Lebo,
Li Ding, Cynthia Chang,
Paulo Pinheiro da Silva, and Deborah L. McGuinness

***Abstract*—The widespread availability of Open Government Data is exposing significant challenges to trusting its unplanned applications. As data are accumulated, transformed, and presented through a chain of independent third parties, there is a growing need for sophisticated models of provenance. Although significant progress has been made in describing data derivation, it has been limited by its inability to distinguish transformations that change content from transformations that merely change representation. We have found that Functional Requirements for Bibliographic Resources (FRBR) can, when paired with a derivational provenance model and cryptographic digest algorithms, successfully represent web resource accession, distinguish between transformations of content and format, and facilitate veracity. We show how FRBR concepts, cryptographic digests, and the World Wide Web Consortium's emerging provenance standard can be used to provide an automated method to coordinate the many, parallel identities of information resources.**

***Index Terms*—open government data, identity, provenance**

I. Introduction

OPEN Government Data (OGD) is a new and rapidly growing phenomenon. Catalyzed in 2009 by countries such as the United States and United Kingdom, governments from local to national levels are publishing their data for public use. [1] These data are available for personal or commercial use and offer the potential to increase the quality of life for communities, businesses, and government alike. Such benefits could include helping citizens understand pollutants near their home, crimes in their neighborhood, public works, natural disasters, and political activities. Further, while individual datasets are interesting on their own, there is a hope and expectation that combining disparate datasets will lead to even more insight and value; the whole should be greater than the sum of its parts.

Unfortunately, combining datasets is more difficult than simply providing each as data files on a web site. A number of social and technical challenges remain. Simply "releasing" data, even with good documentation, does not make it useful. First, consistent or automated ways to discover, access, or obtain new datasets are not ubiquitous. Second, after a dataset is obtained, it is often difficult to quickly and easily merge it with others because it is likely to differ in formatting (zip, csv, xml), modeling paradigms (tabular, relational, hierarchical), use domain-specific terminology, use shortcuts and abbreviations that are difficult to interpret, and refer to entities in differing ways (e.g., "POTUS" and "Barack Obama"). These low-level challenges need to be addressed for each dataset before one can begin to address more interesting high-level questions. Challenges are further compounded by the fact that groups around the world are undertaking similar uncoordinated activities to discover, collect, interpret, analyze, use, publish, and display results derived from the same data sources.

Linked Open Government Data (LOGD), [2] the integration of OGD using semantic web and Linked Data principles, has the potential to meet the unmet expectations for a valuable, combined whole of disparate government datasets. According to Linked Data design principles, the Resource Description Framework (RDF) is used to associate data elements within each dataset. When data elements are named using web-accessible Uniform Resource Identifiers (URIs), they not only get a global name but also provide a direct way to request more information about that entity. For example, when observing a data value "ID", one may need to seek documentation, contact another person for help, or make an educated guess at its meaning. Instead, by using a URI such as <http://logd.tw.rpi.edu/id/us/state/Idaho>, the data element leads to documentation and supplemental description when its identifier is requested from the web using HTTP.

Because relationships are also named with URIs, they offer the same benefits. For example, an organization or company may be based near Idaho, so an RDF triple such as <http://tw.rpi.edu/orgpedia/page/company/0000321150> <http://xmlns.com/foaf/0.1/based_near> <http://logd.tw.rpi.edu/id/us/state/Idaho> leads to supplemental information about not only the IDAHO POWER CO and Idaho, but also how they relate – simply by requesting any of the three URIs from the web.

As part of the LOGD community, the Tetherless World Constellation at Rensselaer Polytechnic Institute has been developing tools and exploring how to apply Linked Data principles to integrate and use government data. The project's primary tool, *csv2rdf4lod*[1], [2] embodies a URI design and data transformation methodology tailored to collect, retrieve, convert, enhance, and publish original government data sources as RDF while maintaining provenance. Developed over the past two years as dozens of team members have processed thousands of datasets, the design enables us to accumulate and derive additional value

James P. McCusker, Timothy Lebo, Li Ding, Cynthia Chang, and Deborah L. McGuinness are with the Tetherless World Constellation, Rennsselaer Polytechnic Institute, Troy, NY, USA. Paulo Pinheiro da Silva is with CyberShARE Center, University of Texas at El Paso, El Paso, TX, USA

[1]https://github.com/timrdf/csv2rdf4lod-automation/wiki

at any stage while ensuring data are backward-compatible and annotated with provenance.

The *csv2rdf4lod* tool uses an RDF vocabulary[2] that describes how tabular structures should be interpreted to create well-structured RDF representations that reuse existing vocabularies and explicitly connect entities common across multiple datasets. Using RDF-encoded, declarative "enhancement parameters" reduces the need for custom software and thus reduces both the likelihood for human error and the time required for a third party to familiarize with the specified enhancement. An important benefit of enhancement parameters is the ability to automatically reproduce the conversion of all tabular datasets using the same uniform terminology, which avoids creating conversion software tailored to each dataset. Further, using RDF to encode the transformation allows anyone to find data products according to how they were transformed by using SPARQL query. This unique capability is naturally available in semantic web technologies.

Although creating Linked Data from government data reduces integration costs and increases the potential for insights and value, it implicitly raises challenges for those choosing to use Linked Data instead of the original form. One issue is that the Linked Data version is often hosted by third parties instead of maintaining the original host. Additionally, the third party is providing a *transformed* version of the reputable data originally provided by the government. What assurances does a consumer have that the data from a third party is *just as good* as that from the government? Do the benefits of integrated and comprehensible datasets provided by the third party outweigh the risks that they may contain mistakes, or, worse, malicious intent? If the same original government dataset is integrated by two different third parties, which should a consumer use?

Even when a variety of provenance information is available, it is difficult to develop trust with data consumers. First, transformations introduce risk to the veracity of the output: if a third party transforms a trusted government data file to produce a result, is that transformed data just as reputable as the original data? Second, humans need to be involved in managing the data. However, people think in terms of high-level data transformations, but most existing provenance representations record low-level operations. We will therefore discuss how information is structured from concrete to abstract and how that can be used to provide transparency into how OGD is used.

A. Use case: Trusting Integrated Data

We describe a simple use case to provide an example of the challenges that Linked OGD consumers face. Although simple, it is prototypical of what we have encountered. The use case also serves as a basis for demonstrating and evaluating our technical approach. Figure 1 illustrates the four actors and seven resources[3] involved:

A Government: provides a single CSV file at a URL[4]. Two other URLs (URL 1 and URL 2) point (i.e., redirect) to the CSV file. Two Data Integrators (W and E) independently retrieve URLs 1[5] and 2[6], respectively, and store results locally for processing and re-publishing.

Integrator W: rehosts their CSV copy on their own site.

Integrator E: applies three transformations before hosting the results on their own site. Each transformation produces a different RDF file. The first (file://raw.ttl) is derived from the CSV using a naive, domain-independent interpretation. The second (file://raw.ttl.rdf) is derived from the first by re-serializing the RDF model from Turtle syntax to RDF/XML syntax. The third (file://e1.ttl) is derived from the CSV using enhancement parameters developed by a human curator familiar with the original content and Linked Data design.

A Data Consumer: is faced with the decision to use any of the seven data files available: either the two URLs provided by the government, or one of the five third-party integrator results.

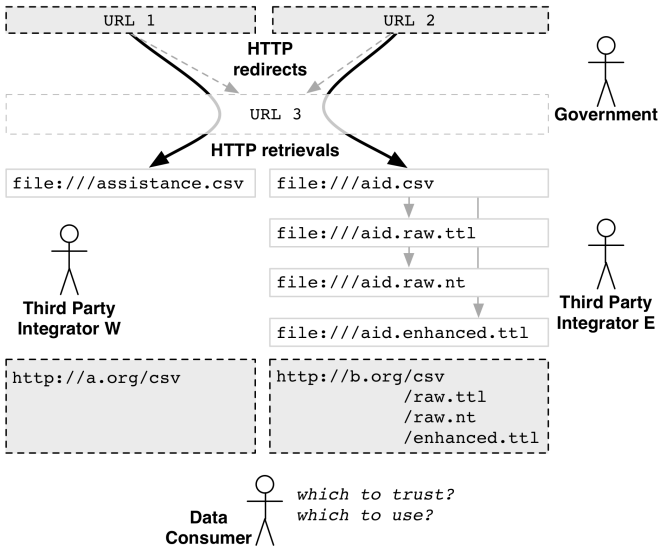

Figure 1. A simple use case where a data consumer must choose between the government's original data or one of five data files offered by third parties.

The challenges center around under-described, uncoordinated proliferation. Understanding the relationships between the choices improves outcomes. While knowing that a file "came from another" is useful, one is much more concerned about whether the *content* of the result is equivalent to what was originally provided. If the data consumer does not trust the integrator, what assurances can the integrator provide to convince the consumer to use their results instead of the reputable form from the government? Third party data integrators need to convince consumers that their results are not only just as good, but *better* than the original; the processed results need to be more discoverable,

[2]http://purl.org/twc/vocab/conversion/

[3]Additional use case information, along with technical details and links to the actual resources, is available at http://purl.org/twc/pub/mccusker2012parallel

[4]http://gbk.eads.usaidallnet.gov/data/files/us_economic_assistance.csv

[5]http://explore.data.gov/download/5gah-bvex/CSV

[6]http://www.data.gov/download/1554/csv

comprehensible, discoverable, and integrated – all while preserving the content and reputability of the original source.

II. Related Work

We cover four kinds of related work: RDF conversion tools, current provenance models, information models from Library Science, and existing content-based cryptographic digests.

A. RDF Conversion Tools

Of the many tools for creating RDF representations, few are mature, in active development, and supported by their developers. For a recent survey that evaluated two dozen leading tools for what they call *knowledge extraction*, see LOD2's report[3]. Google Refine[7] is one prominent tool because of its ease of use. It offers a web interface to open and modify tabular data, while an extension from DERI further permits a curator to construct templates and export RDF results. Although Refine's user interface is easy to use for individual datasets, current design limitations make it difficult to scale to the number of datasets available as OGD that need to be exposed as Linked Data. DERI's export extension also does not provide any reasonable default URI construction, further increasing the amount of human effort required to consistently create well-structured RDF representations for distributed tabular datasets. As with any monotonous human labor, risk of human error is also a concern.

B. Current Provenance Models

Current provenance models describe the provenance of derivation and events relatively successfully. Models like the Open Provenance Model (OPM) [4], Proof Markup Language (PML) [5], and the emerging World Wide Web Consortium (W3C) standard for provenance, PROV,[8] describe the derivational history of information and other entities. These provenance models tend to describe derivation links as edges between entities. OPM and PROV also describe events as additional nodes in the same graph. We call these sorts of events and links derivational provenance, since both record what happened and where things came from.

C. Models from Library Science

Functional Requirements for Bibliographic Records (FRBR) [6] is a model developed by the library science community to describe the world of different bibliographic resources, where an author's work can assume many forms such as a paperback book, eBook, or audiobook. After almost twenty years of development, the Library of Congress, the National Agricultural Library, and the National Library of Medicine have announced their intention to adopt systems based on the FRBR model[9]. Figure 2 illustrates and describes an example to provide an introduction. We call an Item's connection to its Manifestation, Expression, and Work a "FRBR stack". Core FRBR has no actual derivational provenance model, but the OWL representation[10] has minimal properties for creating derivational links within each layer.

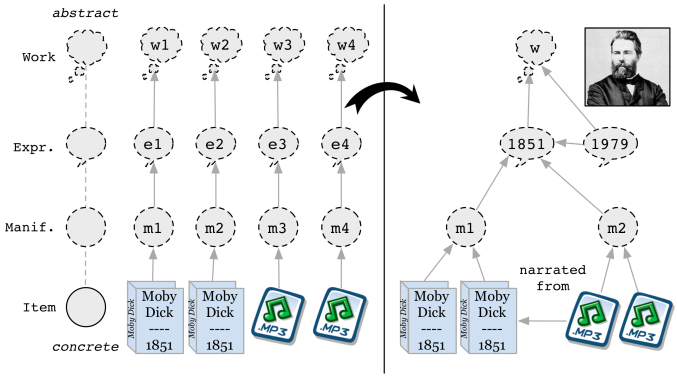

Figure 2. FRBR uses four levels of abstraction (Item, Manifestation, Expression, and Work) to distinguish among parallel aspects of an author's work. Two "identical copies" of a book are distinct Items because they occupy different physical space, but they share the same Manifestation because they have the same physical structure. Audio recordings share the same Expression with the book because they both convey the same content. When the same conceptual story (Work) is reworded slightly by the author, a new Expression is created and is associated with the same Work as the previous Expression.

D. Existing Content-Based Cryptograhic Digests

As it becomes easier to shift between data formats, the ability to verify that information is the same has become weakened because crytpographic message digests work only at the byte level. Two RDF graphs that assert "George a foaf:Person." can be serialized in any number of ways, none of which changes the content of the graph. RDF graph digest algorithms [7] have been developed that are resilient to assertion ordering and other issues. Additionally, strategies such as canonical serialization have been used for other non-graph representations such as XML[8]. Finally, work in creating content-based digests for moving and still images have resulted in the ability to identify image-based content as being identical across a large number of mechanical transformations. [9]

III. Approach

To support more complete explanations of digital information resources, we looked to representations of other sorts of information resources. Bibliographic resources, such as books, albums, films, magazines, etc. are, at their core, information resources. FRBR's four levels of abstractions also apply easily to electronic information sources. For instance, copies of files (Items) are exemplars of the same Manifestation. For images, conversion from JPEG to PNG results in a new Manifestation, but the actual stored image is the same Expression. Finally, if an image is processed, the raw and post-processed images are then different Expressions of the same Work.

Use of cryptographic digests makes it possible to automatically identify various levels of electronic information resources. Since Manifestations correspond easily to particular

[7]http://code.google.com/p/google-refine/
[8]http://www.w3.org/2011/prov/wiki/WorkingDrafts
[9]http://www.loc.gov/bibliographic-future/rda/
[10]http://purl.org/vocab/frbr/core#

data streams, it is possible (and is the principal application of cryptographic digests) to create a unique, repeatable number – a *message digest* – to identify that data stream. Anyone else who encounters that data stream can compute the same digest. Similarly, digest algorithms have been developed for RDF graphs [7] that result in the same hash regardless of the order or original serialization format. Use of these sorts of algorithms supports reproducibility of content - the same *content digest* identifies the same information. If two message digests (Manifestations) differ, but share the same content digest (Expression), then the content is serialized in alternative representations.

As graph digests are only useful for RDF graphs, we have identified requirements for *content digests* in general that can be used to automatically identify Expressions. Effort must be made to find a format-invariant interpretation of the file contents using a single number, as with message digests. However, the same content does not always need to map to the same digest, for instance, it is acceptable to fall back to message digests if no content digest is available. We have identified some example content digest implementations that provide robust content identity across different manifestations:

Table Digest: Take the graph hash aggregate of every cell where the cell is a tuple (row, column, value). For files with multiple sheets, the tuple would be (sheet, row, column, value).

Relational Digest: For every row in every table in the database, create an aggregate digest of each fieldname/value pair, and then create an aggregate digest of the tuples (table-name, rowdigest). This can be computed on the fly in databases using triggers for insert, update, and delete operations if the aggregate function is addition.

Image Digest: Existing algorithms have been identified. [9].

IV. Methods

Message-level and content-level digest algorithms were implemented in two stand-alone python utilities. The first, fstack.py, produces an RDF description using terms from the Functional Requirements for Information Resources (FRIR) vocabulary,[11] which was created to extend Ian Davis' FRBR-core ontology,[12] [6] Nepomuk's File Ontology, and W3C's draft PROV ontology. The second utility, pcurl.py, produces a similar RDF description for a file retrieved from a URL, but includes information about the URL and its HTTP response.

The csv2rdf4lod-automation data integration toolset was extended to incorporate the results from pcurl.py and fstack.py when retrieving URLs and when converting data files. The use case described in Section I-A was encoded in a shell script that implemented retrieval, conversion, reformatting, and enhancement.

V. Results

The results of our investigations manifest in the FRIR representations of the files manipulated within the use case. While the original data files are available, we display diagrams here that were automatically constructed from the results. Some abbreviations were made for presentation purposes, including shortening the cryptographic digests in the URIs naming the Items, Manifestations, Expressions, and Works. The *consolidation of higher-level endeavors* is the principal characteristic to consider when observing FRBR stacks of files and their manipulations; when higher levels are consolidated, more information is known about the more concrete forms and whether or not they can be used for a particular application. Figures 3, 4, 5, and 6 show successful implementation of the use case. For further disussion, see our online appendix.[13]

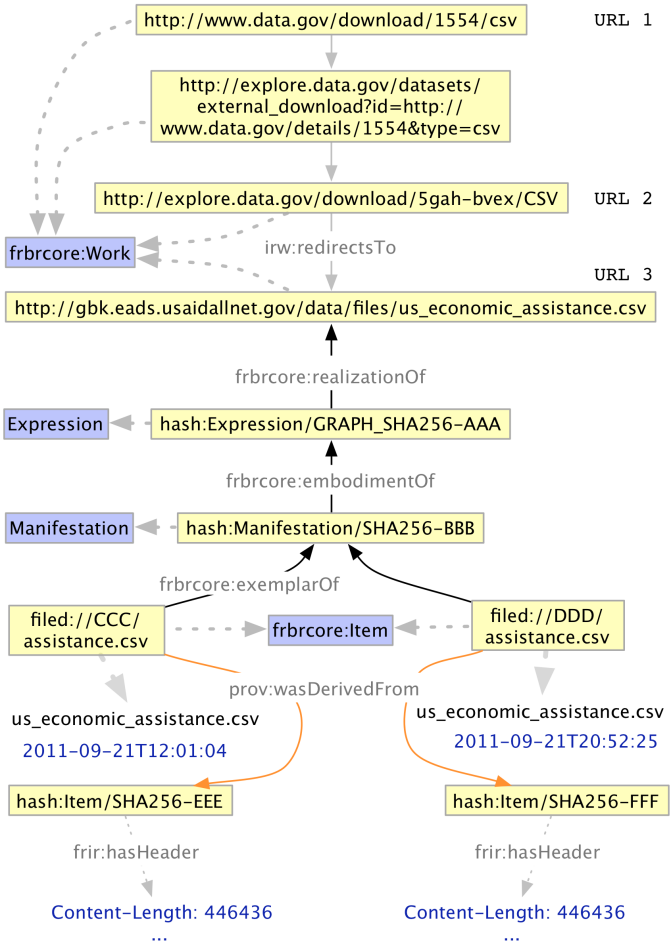

Figure 3. FRBR provenance when Data Integrators E and W retreive two different URLs. The relations among the requested URLs becomes apparent: URL 1 (eventually) redirects to URL 2, which redirects to URL 3. Although retrieved independently, the files share the same Manifestation and Expression because the message digest and content digest were used to name them, respectively.

VI. Discussion

The ability to tell what kinds of transformations are recorded in provenance makes it simpler to show relevant provenance information to users. Additionally, using cryptographic content digests as identifiers makes it simple to verify the identity of content and prove that the same information is used in multiple settings out of band. The uncertainty of not knowing what is contained in each file is managed through the automatic

[11] http://purl.org/twc/ontology/frir.owl
[12] http://purl.org/vocab/frbr/core
[13] http://purl.org/twc/pub/mccusker2012parallel

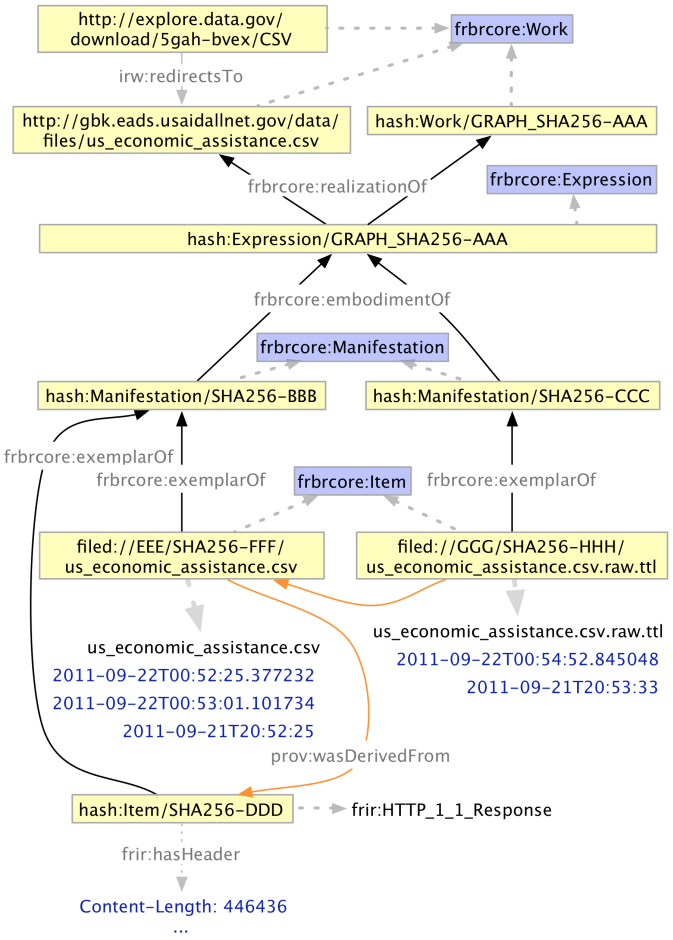

Figure 4. FRBR provenance when Data Integrator E converts the CSV to raw RDF. Although the files' Manifestations differ, the Expression is the same. By this, we know that no new content was created (or lost) in the conversion.

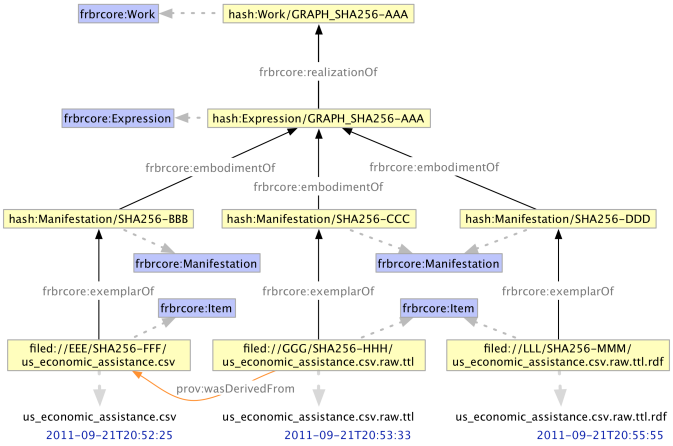

Figure 5. FRBR provenance of the CSV, raw RDF, and a conversion of the raw RDF into RDF/XML. Although the RDF/XML is not stated to be derived from the original, their common Expression permits us to view them as content-equivalent.

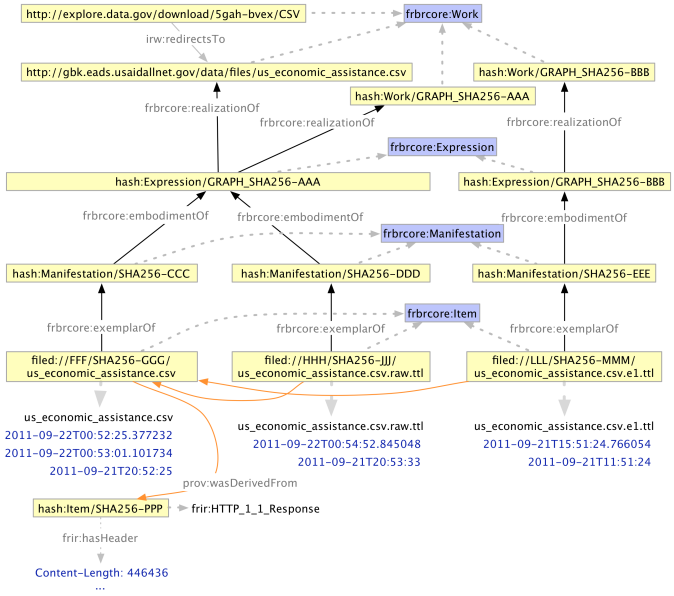

Figure 6. FRBR provenance applying enhancement parameters to the CSV's conversion to RDF. Although the raw RDF's Expression was recognized as tabular and mapped to the same Expression as the CSV, the enhanced RDF is a new graph. This new content structure results in a digest derived from the RDF Abstract Model instead of a table. The new Expression is associated to a new derived Work.

combination of FRBR entities using cryptographic digests. The expansion of these digests will allow for further validation of content across representations, making sure that the content is the most important aspect of data transmission, not the format.

For our data conversion use case, consumers can verify that the raw RDF conversion we provide has the same content as the file retrieved from the government. When they use an enhanced RDF conversion, we can tell them that it was derived from something they trust. Additionally, our extension of FRBR to electronic information resources and use of content and message digests to identify these resources should make it much easier for digital libraries to manage resources that are both electronic and physical. [10]

Because of the fine-grained identity assertions that can be made, independent third parties can provide assurances to data consumers that they are producing and providing data that is just as good as the original data, and in cases where enhancements have taken place, is better than the original. Two independently generated raw conversions of government datasets can be trusted directly because of matched content digests, and trust of the enhancement can be earned by inspection of the results and conversion parameters.

We believe that these sorts of abstractive relationships among entities are important for accurately expressing the provenance of information resources. Currently, the emerging W3C PROV standard has a property reserved for abstractive relationships. It is our view that, as it provides significant value in expressing provenance for information resources, PROV should include FRBR relationships and classes in the core ontology. Barring that, an extension that includes FRBR should be recommended so it can be applied to information resources.

A. *Future Work*

We would like to investigate the use of FRBR to handle composite workflows and for providing high-level visualiza-

tions of workflow history. Higher granularity workflows work at more concrete levels, while lower granularity works at more abstract levels, and similarity of traces may be able to be determined through analysis of workflow at higher levels using FRBR. We plan to deploy this infrastructure to an end-to-end application using the LOGD US-UK foreign aid example. While the use case presented only uses one dataset, it should be possible to show how datasets can be combined from multiple sources. We can also provide veracity of enhanced conversions by supplying digests of the original data, enhancement parameters, and the resulting enhanced data. Users can then reproduce the original conversion and verify it via content digests.

Additional work is needed in new types of content hashes for other types of media. We have covered knowledge graphs, spreadsheets, databases, and moving and still images. Other media types, including audio, video with audio, and text, need to be explored to determine if they can be given content digests. Content digests for these types would make nearly all information resources identifiable by their content. As part of the *csv2rdf4lod* project, we are collecting and developing MIME type-based content digest algorithms and welcome external contributions.

VII. Conclusions

As part of the LOGD project, we perform aggregation and curation of OGD, by applying Linked Data principles to generate LOGD. With OGD, combining datasets with the semantic web gives value to that data. However, the need for data consumers to trust what has been done to the data requires an accurate picture of what content has been created and how it has been modified. We developed a use case that expresses these needs, and showed that using FRBR to build multiple levels of abstraction of information resources, when paired with content-based cryptographic digests, allows for easy identification and validation of information resource content. The use of these digests to identify Expressions makes it possible for data consumers to trust third parties with management of data by making that management transparent at a level that is relevant to the consumer. This use of multiple levels of identity, especially content-based identity, makes it possible for data consumers to trust what modifications, if any, that have been made to the data they use. As our LOGD system is a form of digital library, our experiences with improving trust and transparency can possibly be applied to that domain as well. This paper provides a way to assure that consumers are getting the "same stuff" that they asked for.

References

[1] D. Robinson, H. Yu, W. Zeller, and E. Felten, "Government data and the invisible hand," *Yale Journal of Law & Technology, Vol. 11, p. 160, 2009*, 2009.

[2] T. Lebo, J. S. Erickson, L. Ding, A. Graves, G. T. Williams, D. DiFranzo, X. Li, J. Michaelis, J. G. Zheng, J. Flores, Z. Shangguan, D. L. McGuinness, and J. Hendler, "Producing and using linked open government data in the twc logd portal (to appear)," in *Linking Government Data* (D. Wood, ed.), New York, NY: Springer, 2011.

[3] S. H. et al., "Report on knowledge extraction from structured sources," Tech. Rep. Deliverable 3.1.1, LOD2 - Creating Knowledge out of Interlinked Data, March 2011.

[4] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, *et al.*, "The open provenance model core specification (v1. 1)," *Future Generation Computer Systems*, 2010.

[5] D. McGuinness, L. Ding, P. Pinheiro Da Silva, and C. Chang, "Pml 2: A modular explanation interlingua," in *Proceedings of AAAI*, vol. 7, 2007.

[6] OâĂŹNeill, E.T., "FRBR: Functional Requirements for Bibliographic Records," *Library resources & technical services*, vol. 46, no. 4, pp. 150–159, 2002.

[7] C. Sayers and A. Karp, "Computing the digest of an rdf graph," *Mobile and Media Systems Laboratory, HP Laboratories, Palo Alto, USA, Tech. Rep. HPL-2003-235*, vol. 1, 2004.

[8] H. Maruyama, K. Tamura, and N. Uramoto, "Digest values for dom (domhash)," *Network Working Group, http://www.ietf.org/rfc/rfc2803.txt*, 2004.

[9] F. Lefèbvre, J. Czyz, and B. Macq, "A robust soft hash algorithm for digital image signature," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 2, pp. II–495, IEEE, 2003.

[10] B. Wilson, F. Shaw, L. Vaughn, C. Awre, I. Dolphin, G. Hanganu, T. Brett, C. Ingram, C. Consultancy, M. Custard, *et al.*, "Hierarchical catalog records: Implementing a frbr catalog," *D-Lib Magazine*, 2005.