Parallel Identities for Managing Open Government Data

James P. McCusker, Timothy Lebo, 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 Records (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 that can be used by data consumers to make informed decisions about which data product to use for their application.**

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 including the United States and the 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 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 inherently useful. First, consistent or automated ways to discover, access, and obtain new datasets are not ubiquitous. Second, once a dataset is obtained, it is often difficult to quickly and easily merge it with others because it often differs in formatting (zip, csv, xml) and modeling paradigms (tabular, relational, hierarchical), uses domain-specific terminology, uses shortcuts and abbreviations that are difficult to interpret, and refers 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 explore 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, publish, and display results derived from the same 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[1] when its identifier is requested from the web using HTTP.

Because relationships are also named with URIs, they offer the same benefits. For example, if an organization or company is based near Idaho, 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. By reapplying this Linked Data approach to additional datasets, reused vocabulary and interconnected entities provide an explicit basis for a whole that *is* greater than the sum of its parts. This ability distinguishes RDF from alternative representations such as CSV, XML, and JSON that do not provide intrinsic means to achieve such cross-dataset integration.

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

James P. McCusker, Timothy Lebo, 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] e.g., that it is a state of the United States and was admitted in 1890.

tool, *csv2rdf4lod*[2], [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 of the operations it performs. 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 at any stage while ensuring data are backward-compatible and annotated with provenance. These efforts have resulted in nearly ten billion RDF triples about a multitude of government topics.

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 being maintained on 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?

However, simply providing provenance to data consumers does not mean they will understand it. We claim that a *disparity of abstraction* is the principle barrier for data consumers to trust provenance-annotated data. While humans phrase management of data in terms of high-level abstractions, most existing provenance representations exhaustively record low-level details. Instead of forcing a choice between high-level and low-level provenance, we propose to unify them with four parallel levels that span from abstract to concrete. Any of the four parallel identities of an information resource can then be considered at the level appropriate for a particular task. We describe how structuring aspects of an information resource across levels of abstraction minimizes the need for exhaustive provenance capture and how the resulting formal model can be used to increase 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 many situations that 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[5] and URL 2[6]) point (i.e., redirect) to the CSV file. Two Data Integrators (W and E) independently retrieve URLs 1 and 2, 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, *aid.raw.ttl* is derived from the CSV using a naive, domain-independent interpretation. The second, *aid.raw.rdf* is derived from the first by re-serializing the RDF model from Turtle syntax to RDF/XML syntax. The third, *aid.enhanced.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 offered by the third-parties.

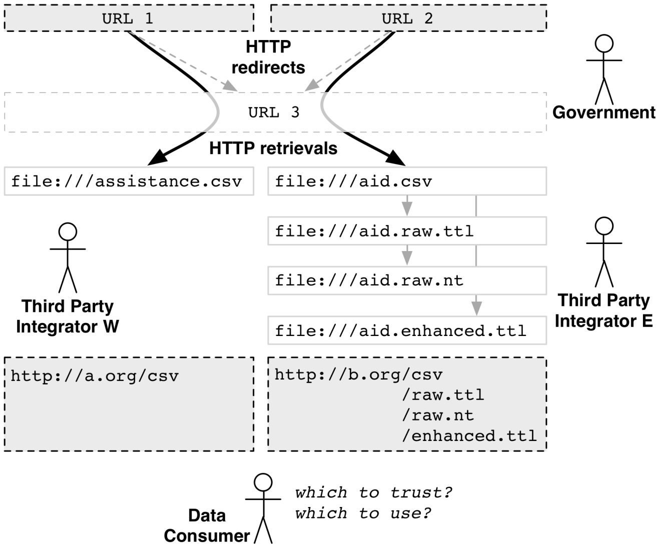

Figure 1. A simple use case where a data consumer must choose between the government's original data or one of five data files offered by third parties.

The consumerÕs challenges center around under-described, un-coordinated proliferation that is characteristic of the Web. Understanding the relationships among the choices can lead the consumer to a more informed and confident decision. Without knowing the nature of the transformations that led to each data file, a cautious consumer must assume their results are different until shown otherwise. If a third party offers a result whose *content* is equivalent to that offered by the original source, the format and transformations leading to it are intrinsically satisfactory. 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, comprehensible, discoverable, and integrated – all while preserving the content and reputability of the original.

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.

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

[3] Additional use case information including 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

A. RDF Conversion Tools

LOD2's recent report [3] surveys two dozen leading tools[7] for *knowledge extraction*. Google Refine[8] is prominent and offers a web interface to modify tabular data, while an extension permits the construction of RDF export templates. Although easy to use for small, individual datasets, Refine makes it difficult to scale to the overwhelming number of OGD datasets that need to be exposed as Linked Data. The export extension also does not provide reasonable default URI creation, which increases the amount of human effort required to consistently name instances across datasets.

The conversion tool used in this paper, *csv2rdf4lod*, uses an RDF vocabulary[9] to describe how to interpret spreadsheets into well-structured RDF representations that reuse existing vocabularies and explicitly connect entities across datasets. Unlike the R2RML language in development by the W3C, the vocabulary that *csv2rdf4lod* uses does not assume an underlying relational database. Instead, it borrows design from RDFS and addresses a wider variety of tabular encodings, including n-ary and statistical data. Using declarative "enhancement parameters" [2] reduces the need for custom software and thus reduces the likelihood for human error and the time required for a third party to familiarize with the enhancement. Further, declarative parameters enable others to automatically reproduce conversions using the same uniform terminology.

each dataset.

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,[10] 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 intent to adopt systems based on the FRBR model[11]. Figure 2 illustrates an example of how FRBR can be applied to organize the different aspects bibliographic resources. FRBR uses four levels of abstraction 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 revised, a new Expression is created and is associated with the same Work as the previous Expression. We call an Item's connection to its Manifestation, Expression, and Work a "FRBR stack". Although core FRBR has no derivational provenance model, its OWL representation[12] provides some minimal properties to create derivational links within and across levels of abstraction.

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 cryptographic message digests work only at the bitstream level. For example, 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. Strategies such as canonical serialization have been used for other non-graph representations, including dataset publication using the Universal Numerical Fingerprint (UNF) [8]. Finally, work in creating content-based digests for images and movies can identify image-based content as across a large number of mechanical transformations. [9]

III. Approach

By applying FRBR's four levels of abstraction to *digital* information resources and naming their four parallel aspects using cryptographic digests, we can support useful explanations of manipulated data products. FRBR's bibliographic resources such as books, albums, films, and magazines are, at their core, information resources. FRBR's four levels of abstraction also naturally apply to electronic information sources. Copies of files (Items) are exemplars of the same Manifestation (byte sequence). Similarly, an Excel file created from a CSV will have a different Manifestation, but maintain the same Expression because they both store the same data. Finally, if the data is modified, the original and resulting spreadsheets have different Expressions (visual content) of the same abstract Work.

Figure 2 illustrates how this digital FRBR approach can be applied to organize the use case's data products according to their four parallel identities. Any file (Item) retrieved from the two government URLs or their rehosted locations will result in the same bitstream, and thus share an identical Manifestation. The naive transformation producing *raw.ttl* does not change the tabular content of the original CSV, so they share identical Expressions while differing in Manifestations (bitstream). The reserialized RDF model in *raw.rdf* again changes Manifestation, but retains an Expression identical to both *raw.ttl* and

[7] http://purl.org/twc/page/tabular-rdf-converters lists more.

[8] http://code.google.com/p/google-refine/

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

[10] http://www.w3.org/2011/prov/wiki/WorkingDrafts

[11] http://www.loc.gov/bibliographic-future/rda/

[12] http://purl.org/vocab/frbr/core#

the CSV. The Item *enhanced.ttl* does not preserve an identical (tabular) Expression as the others because it was restructured during curation and can contain more or less content. The enhancement's Work is also distinct because its content was created by Integrator E instead of the government.

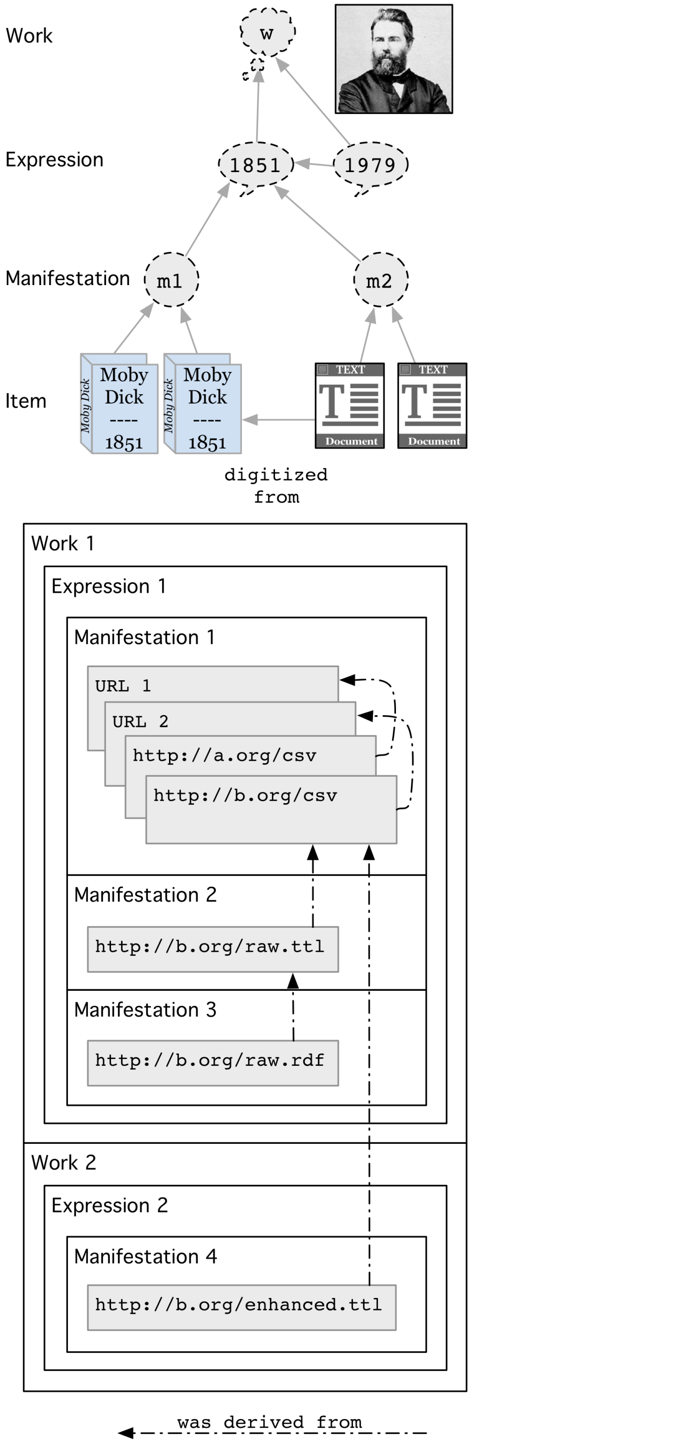

Figure 2. The data products from the use case in Figure 1 can be organized according to their four parallel identities established by FRBR's Work, Expression, Manifestation, and Item.

Cryptographic digests makes it possible to automatically identify electronic information resources in a number of ways. First, since the physical structure of Manifestations correspond to the sequences in data streams, it is possible to create a unique, repeatable number – a *message digest* – to identify that data stream. This is the principal application of cryptographic digests; anyone else who encounters that data stream can compute the same digest and know that they have received the same sequence. Similarly, digest algorithms have been developed for specific content types that apply independently of its serialization. Digests for RDF graphs [7] produce the same hash regardless of statement order or serialization format. Digest algorithms can be used to recognize reproduced 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.

Although message digests apply to any digital file, content digests only apply to specific types of content. Effort must be made to find a format-invariant interpretation of the file contents identified by a single number. Because graph digests are only useful for RDF graphs, we have identified requirements for other *content digests* that can be used to automatically identify Expressions. While mapping the same content to the same digest is ideal, mapping it to different digests is also acceptable as an approximation. One may conservatively fall back to identifying an Expression using the message digest of the Manifestation that embodies it in situations where no content digest is available. In addition to the content digest types discussed above, we have identified a digest algorithm for raw spreadsheet tables. Simply 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).

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,[13] which was created to extend Ian Davis' FRBR-core ontology,[14] [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. We created a script that performs the retrieval, conversion, reformatting, and enhancement in the use case described in Section I-A. It produces four files to describe different combinations of the events that took place. The first compares the FRBR stacks created when Integrators W and E retrieve URLs 1 and 2, respectively. The second compares the FRBR stacks created when Integrator E converts the CSV to Turtle with a naive, domain-independent interpretation. The third compares the FRBR stacks of the CSV, the Turtle derived from the CSV, and the RDF/XML derived from the Turtle – all created by

[13] http://purl.org/twc/ontology/frir.owl

[14] http://purl.org/vocab/frbr/core

Integrator E. The fourth compares the FRBR stacks of the same CSV and Turtle to the FRBR stack of the enhanced Turtle – all created by Integrator E. Each of the four comparison files were inspected for common parallel identities between the FRBR stacks that were independently constructed during the use case. Source code, results, and further details about the apparatus are available at our online appendix[15].

V. Results

Figure 3 illustrates the first[16] of the four comparison files created, where Integrators W and E request different URLs and receive distinct files with the same Manifestation (message digest). The result shown is a union of two independently-asserted FRBR stacks. Integrator W mentions URL 1, while Integrator E does not. Similarly for Integrator E and URL 2. Both mention URL 3 and identify the same Work, Expression, and Manifestation, which correspond to *URL 1*, *Expression 1*, and *Manifestation 1* in our objective organization in Figure 2.

Figure 3 was automatically constructed from the comparison file. 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.

In the second comparison file, the message digest used to identify the Manifestations of the CSV and Turtle files differ. This tells us that the physical structure of the files differ. Because both files convey tabular content, the content digest used to identify their Expressions are identical. Although we omit the result here, the structure can be seen in Figure 2 with *http://b.org/csv* and *http://b.org/raw.ttl*. In the third comparison file, where the Turtle derived from the CSV is reserialized to RDF/XML, the message digests used to identify all three Manifestations differ, yet they all share the same Expression because the tabular content digest recognized identical tabular content. Again, this structure can be seen Figure 2 with *http://b.org/csv*, *http://b.org/raw.ttl*, and *http://b.org/raw.rdf*. In the fourth comparison file, the tabular content digest did not apply to *http://b.org/enhanced.ttl*, because the curated Turtle did not exhibit a tabular structure like *http://b.org/csv* and *http://b.org/raw.ttl*. Instead, the RDF graph digest was used to identify the enhancementÕs Expression. This structure, too, is omitted for space but can be seen in Figure 2.

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 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.

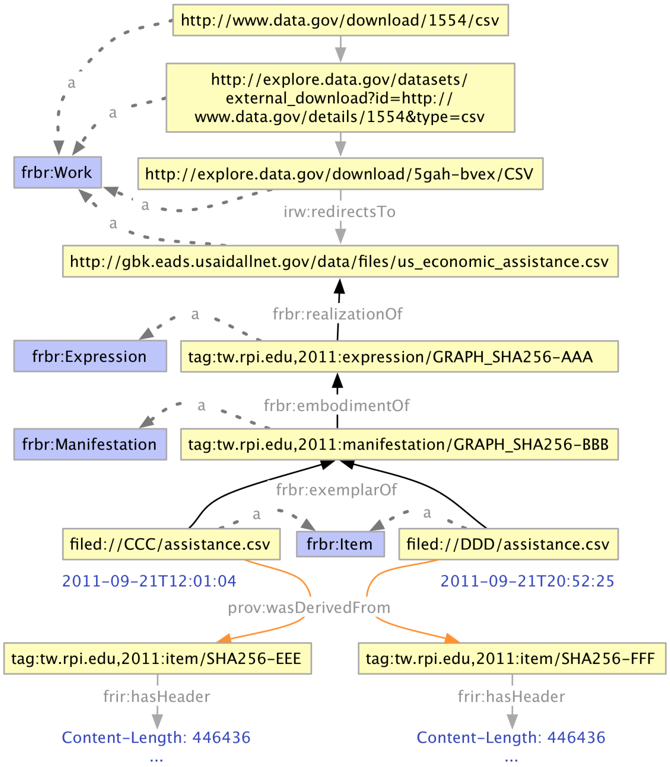
http://www.data.gov/download/1554/csv
http://explore.data.gov/datasets/ external_download?id=http:// www.data.gov/details/1554&type=csv
frbr:Work
http://explore.data.gov/download/5gah-bvex/CSV
a
a
a
a
irw:redirectsTo
http://gbk.eads.usaidallnet.gov/data/files/us_economic_assistance.csv
frbr:realizationOf
a
frbr:Expression
tag:tw.rpi.edu,2011:expression/GRAPH_SHA256-AAA
frbr:embodimentOf
a
frbr:Manifestation
tag:tw.rpi.edu,2011:manifestation/GRAPH_SHA256-BBB
frbr:exemplarOf
a
a
filed://CCC/assistance.csv
frbr:Item
filed://DDD/assistance.csv
2011-09-21T12:01:04
2011-09-21T20:52:25
prov:wasDerivedFrom
tag:tw.rpi.edu,2011:item/SHA256-EEE
tag:tw.rpi.edu,2011:item/SHA256-FFF
frir:hasHeader
frir:hasHeader
Content-Length: 446436
...
Content-Length: 446436
...

Figure 3. FRBR provenance when Data Integrators E and W retrieve 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. The unlabeled dashed lines are rdf:type triples.

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 W3C

[15] http://purl.org/twc/pub/mccusker2012parallel

[16] Illustrations of all comparisons from the use case are available at our online appendix.

Provenance Working Group, in its work to develop a provenance standard for the web, is including a property to express 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 to encourage re-use. 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 visualizations 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, 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] O'Neill, 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] M. Altman, "A fingerprint method for scientific data verification," *Advances in Computer and Information Sciences and Engineering*, pp. 311–316, 2008.

[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.