The Coming Web of Genealogical Data

Josh Hansen
FamilySearch
joshhansen@ldschurch.org

Abstract

The long-held dream of a single, global family tree is now within reach. What if users of different family tree websites could link to people, places, and events on other sites? What if everybody was, in fact, working on one tree? For various reasons this goal remains unrealized, though not for lack of effort. Fortunately, recent developments in the Semantic Web provide a sound technical and organizational foundation for achievement of this vision. In this paper I describe how Semantic Web technologies and the Linked Data philosophy have led to the emergence of a global Web of Data—and why stewards of genealogical data should care. I then suggest an integrated Semantic Web vocabulary for representation of genealogical data. Finally, I explain some of the benefits a Linked Data approach to genealogy can provide to users, and one way your website can join the Data Web.

1 Introduction

Family history technology is in the middle of an ongoing process of integrating previously separate data silos. By means of websites such as New FamilySearch, Ancestry World Tree, and WeRelate, data previously kept isolated on individual computers is now increasingly stored in common spaces available to many users simultaneously. This reduces duplication of effort, accelerates the pace of research, and improves the quality of the resultant family tree. However, even the largest of these sites are far from comprehensive in their coverage. Moreover, these sites remain largely isolated from each other, resulting in many of the same problems of redundancy as before, but on a grander scale.

Fortunately, Semantic Web technologies and Linked Data principles provide a clear technical solution to the challenge of integrating these data silos into a truly global family tree. In this paper I describe the core technologies and practices that constitute the Semantic Web in its current form. I describe how Semantic Web technologies and the Linked Data philosophy have led to the emergence of a global Web of Data—and why stewards of genealogical data should care. I then suggest an integrated Semantic Web vocabulary for representation of genealogical data. Finally, I explain some of the benefits a Linked Data approach to genealogy can provide to users, and describe a method by which your website can join the Data Web.

2 The Semantic Web

The Semantic Web consists of a set of technologies enabling a global data space. Semantic Web data can be of any type. It can be published by anybody and represent multiple, even contradictory, viewpoints. Entities on this Data Web are interconnected via links (as the classical Web is), facilitating discovery of related data. In short, the Semantic Web does for *data* what the ordinary Web has done for *documents.*(Heath and Bizer 2011, ch. 3)

The foundational technology of the Semantic Web is the Resource Description Framework, or RDF. RDF expresses data as *properties* of *things*. In RDF, *things* (including property types) are represented by *URLs*, more prop-

erly termed *URIs*.[1] For example, the actor James Stewart could be represented by the URI *<http://example.com/actors/JamesStewart>* and his movie "Rear Window" could be represented as *<http://example.com/movies/RearWindow>*. A property of interest for an actor might be *<http://example.com/properties/actedIn>*, indicating a movie that the actor acted in.

RDF makes assertions about the properties of various things using statements in the form

subject predicate object .

where $subject$ is the URI of the thing being described, $predicate$ is the URI of the type of property being described, and $object$ is the URI or literal value (string, integer, date, etc.) of the value of that property. So, to assert that James Stewart acted in "Rear Window", we would use the following statement:

<http://example.com/actors/JamesStewart>
<http://example.com/properties/actedIn>
<http://example.com/movies/RearWindow>
.

James Stewart's date of birth—a literal value rather than a URI—can be asserted as

<http://example.com/actors/JamesStewart>
<http://example.com/properties/birthDate>
"1908-05-20"^^xsd:date .

These statements are commonly known as *triples* and are the fundamental units of meaning on the Semantic Web.(Manola and Miller 2004, ch. 2) A huge proportion of the world's data can be effectively represented as combinations of RDF triples.

3 Linked Data

In addition to foundational technologies such as RDF, deployments of data on the Semantic Web have been facilitated by Tim Berners-Lee's formulation of the *Linked Data* best practices. These practices are summarized by the following four principles:

1. Use URIs as names for things.
2. Use HTTP URIs, so that people can look up those names.
3. When someone looks up a URI, provide useful information, using standards like RDF.
4. Include links to other URIs, so that they can discover more things.

(Heath and Bizer 2011, ch. 2, after Berners-Lee 2006)

Linked Data practices have achieved significant uptake in diverse domains including geography, government, media, life sciences, and user-generated content.(Heath and Bizer 2011, sec. 3.2) This is illustrated in Figure 1, which details a subset of the current web of Linked Data. The momentum towards deployment of datasets as Linked Data continues. For example, a substantial new initiative headed by the Library of Congress aims to bring that library's bibliographic data onto the Semantic Web. The arguments justifying this decision have direct relevance for custodians of family history data:

> The protocols and ideas behind Linked Data are natural exchange mechanisms for the Web that have found substantial resonance even beyond the cultural heritage sector. Likewise, it is expected that the use of RDF... will enable the integration of library data and other cultural heritage data on the Web for *more expansive user access to information*....
>
> Embracing common exchange techniques (the Web and Linked Data) and broadly adopted data models (RDF) will move the current library-technological environment *away from being a niche market unto itself to one more readily understandable by present and future data creators, data modelers, and software developers.* It is anticipated that all of these considerations, taken together, will result in greater cost savings for libraries. For example, libraries will be able to take advantage of a broader selection of technological solutions and leverage the knowledge and

[1] To be completely accurate, RDF identifies things by means of URIrefs, which are URIs with optional fragment identifiers (e.g. *http://example.com/movies/RearWindow#scene5*). See Klyne and Carroll 2004, section 3.2.

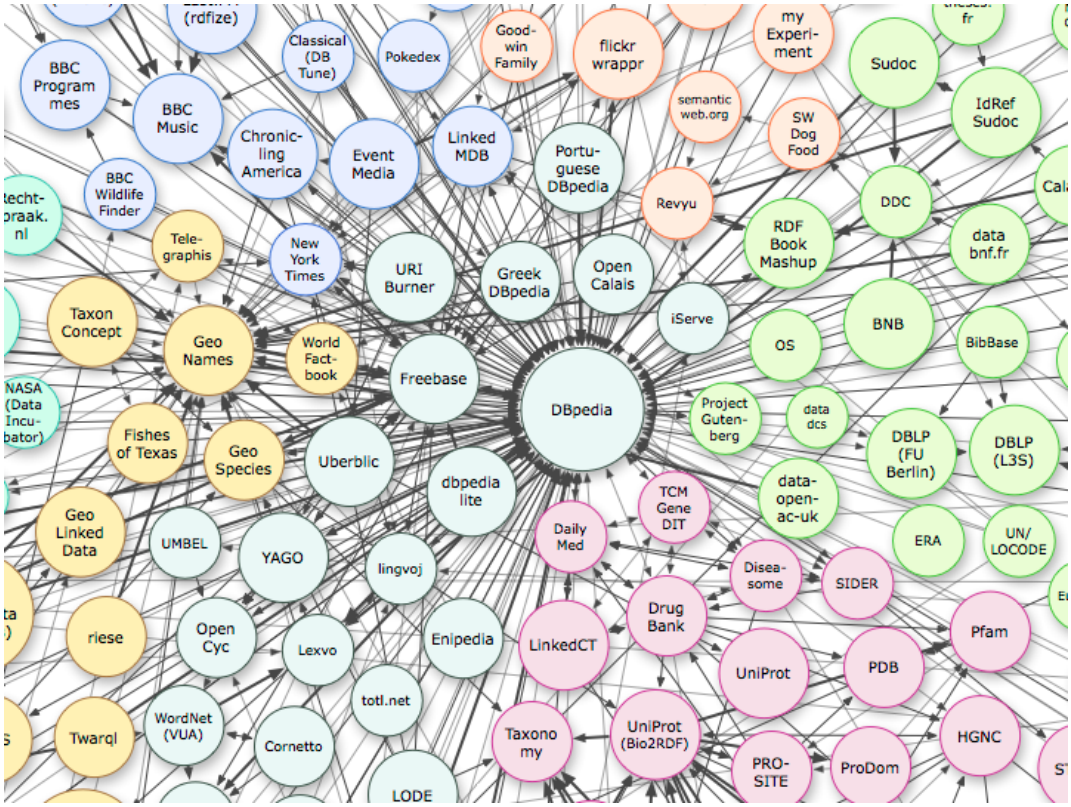

Figure 1: In spite of the obvious utility of a common genealogical data space and the wide breadth of domains already represented on the Semantic Web—including many directly relevant to family history and genealogy, such as FOAF (people and relationships), GeoNames (geographic locations), and Dublin Core (documents and publications)—genealogical data conspicuously remains almost entirely absent from the Data Web. This is in spite of substantial efforts to bring the benefits of the Data Web to genealogical research. Detail from "Linking Open Data cloud diagram", by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

> skills of current and future professionals. Those professionals are, or will be, deeply conversant with more contemporary data creation, data modeling, and software development practices.(Library of Congress 2011, my emphasis)

4 Linked Genealogical Data

The time for integrating genealogical datasets with the Web of Data is *now*.

In recent years, websites such as New FamilySearch, Ancestry World Tree, and WeRelate have created shared family trees available to many users simultaneously. The ability of users to work together on the same tree reduces duplication of effort, accelerates the pace of research, and improves the quality of the resultant family tree. However, no single genealogical service will ever capture all of the market, all of the users, and all of the data. The inability or refusal of service providers to acknowledge this reality does not serve users well.

A key problem with the current arrangement is *synchronization*. For example, if one part of a user's family tree is best covered by WeRelate, and another part is best covered by New FamilySearch, these sections will remain separated from each other without large-scale hand copying of data from one place to another. Even if the means and will to perform such duplication existed, the two services would quickly fall out of sync again as changes on one service or the other were subsequently made. Additionally, the user's awareness of relevant information on other services is limited by the user's ability and willingness to create (and pay for) multiple user accounts, learn multiple user interfaces, and execute the same search multiple times.

The synchronization problem can be resolved by acknowledging that genealogical research is an inherently distributed endeavor. No one data repository will ever contain all of the data, no one researcher will ever do all of the research, and no one website will ever capture all of the users. Yet because of the connectedness of all humans beings, these researchers and users are in fact united in a common effort, and these data yearn to be linked together in a single, global, universal family tree.

Semantic Web technologies and Linked Data principles provide a clear solution to the incredible challenge of connecting these people and datasets. It is a solution that does not depend upon adoption of a single service, but rather will form by enabling existing services to speak the language of the Semantic Web. And though this solution requires some new technologies, such as RDF, it also builds upon many familiar ones, like URIs, HTTP, and HTML.

These techniques have been embraced by numerous other research endeavors, proving the utility and practicality of the approach. They will continue to be embraced by others, broadening the usefulness of the Data Web. As the broader technology and research ecosystems increasingly adopt a Linked Data approach, the relative cost of maintaining genealogical datasets in separate silos and single-purpose data formats will increase. Now is the time for stakeholders in the area of family history technology to come aboard the Data Web and reap the rewards for users, institutions, and the work of genealogy overall.

5 Prior Work

Efforts to include genealogical data in the Semantic Web are many.(Woodbury and Embley 2005, Zandhuis)[23456789] I do not survey these prior attempts in any detail here. Most suffer from use of outdated technology (DAML), failure to reuse widely accepted vocabularies such as FOAF, or outright namespace failure (HTTP 404 responses). John Goodwin's Family Tree[10] is the only extant Linked Genealogical Data deployment I am aware of. It is a good example of what can be accomplished using existing vocabularies, and has strongly influenced my choice of vocabularies in Section 6. My hope is that this work will serve as a standardization of John Goodwin's and others' pioneering efforts and will lead to many more such deployments.

6 A Linked Data Vocabulary for Genealogy

Much of the utility of a Web of Data comes from using the same terms to describe the same kinds of relationships. In order to facilitate the interlinking of genealogical data on the Web, I now introduce the GEN vocabulary. GEN is a simple RDF vocabulary facilitating the deployment of linked genealogical data. It references a set of existing vocabularies that I believe provide a usable, largely integrated framework for linked genealogical data. It also introduces a number of terms not covered in existing vocabularies that I expect to be of use to genealogists. The GEN vocabulary has a preferred namespace prefix of `gen` and a namespace URI of `http://purl.org/gen/0.1#`.

GEN takes an event-centric view of human relationships, i.e. rather than directly stating

JamesStewart mother ElizabethJackson .
JamesStewart father AlexanderStewart .

we interpose an event

JamesStewart birth BirthEvent .
BirthEvent mother ElizabethJackson .
BirthEvent father AlexanderStewart .

Though this approach is not mandatory, it is encouraged as a means of keeping the family tree structures close to (potentially) empirically verifiable events.

In the remainder of this section, I describe the vocabulary's various elements.

[2] `http://www.daml.org/2001/01/gedcom/`
[3] `http://orlando.drc.com/semanticweb/daml/ontology/genealogy/gentology-ont`
[4] `http://www.daml.org/ontologies/107`
[5] `http://jay.askren.net/Projects/SemWeb`
[6] `http://groups.drupal.org/node/19837`
[7] `http://groups.drupal.org/node/156559`
[8] `http://www.owldl.com/ontologies/family.owl`
[9] `http://danbri.org/words/2009/01/18/390`
[10] `http://johngoodwin225.wordpress.com/2009/01/21/genealogy-and-the-semantic-web/`

People and Names

People are modeled using the ubiquitous `foaf:Person` class. Because `foaf:Person` allows for only rudimentary naming, the GEN vocabulary adds an accompanying `gen:Name` class, allowing for an extensible, first-class notion of personal names, as in Woodbury and Embley (2005). For now the only type of name implemented is `gen:FoafName`, which aliases FOAF's `name`, `givenName`, and `familyName` properties. However, by easily allowing for multiple complete names, `gen:Name` provides a foundation for future development of personal name representations for the Semantic Web.

Life Events

The most comprehensive vocabulary for expressing life events is BIO, which includes a fairly complete library of event types, from Birth to Accession to Assassination. The GEN vocabulary includes some of the most common event types from BIO (Birth, Marriage, Death, Baptism, Divorce) as well as some additional event properties such as `birthFather` and `birthMother`.

Families and Family Trees

`gen:Family` represents families of all kinds, allowing members to be linked using traditional, nuclear-family roles (`mother`, `father`, and `child`) as well as `stepMother` and `stepFather`. The generic `familyMember` property allows for other roles to be specified.

In cases where these roles can't fully specify the relationships between family members (e.g. a man is father of three children and step-father of two in the same household) events should be used to express the relationships (i.e. three births to one parent, two births to another, and a marriage.)

`gen:FamilyTree` allows for grouping of people, families, events, sources, and so on and is mainly intended to improve usability when users wish to conceptually group different elements. Both `gen:Family` and `gen:FamilyTree` should also prove useful in transitioning from other data models that include explicit representations of these concepts.

Source Citations

The `gen:Assertion` class allows for source citations in support of any assertion or group of assertions. `Assertion`s can be nested within each other, and each distinct assertion can be attributed to any number of sources. Sources can be of any type, though individuals in class `bibo:Document` or class `frbr:Work` are recommended. In addition to `Assertion`, the GEN vocabulary introduces specific source types relevant to genealogical research. These include `Microfilm`, `Microfiche`, and `GraveInscription`.

Future Revisions

The GEN vocabulary should be considered a first draft in what should be a community process of building consensus around a common vocabulary. I urge members of the family history technology community to examine the GEN vocabulary and assess what improvements might be made in order to more ably represent all use cases and gain the broadest possible adoption. I am particularly interested in discovering inconsistent semantics, omitted terms, unnecessary classes (are events sufficient, or do we need `gen:Family`?), and any other oversights, large or small, so the vocabulary can be improved. The latest draft of the vocabulary can be found at `https://github.com/joshhansen/vocab-gen`. Please send feedback to joshhansen@ldschurch.org.

A Note on URIs

In addition to a vocabulary, an important element of Linked Data deployments is a stable URI scheme for identification of entities being modeled. Each service provider will have to establish a naming system by which URIs are "minted" that will allow for consistent reference to the resources. This is a critical process and should not be overlooked in deployment of Linked Genealogical Data. See Sauermann and Cyganiak (2008) for more information.

7 Benefits to Users

In spite of massive numbers of related records, genealogical datasets held by different organizations—or even in the same organization—are completely isolated from each other. Facts about an individual

		Name	Birth or Christening	Death or Burial	Spouse	Parents
What You Searched For						
		James Maitland Stewart	20 May 1908			Alexander Maitland Stewart Elizabeth Ruth Jackson
Close Matches: 1						
Go to		James Maitland Stewart ★★★★★ @FamilySearch.org	20 May 1908 Indiana,Indiana,Pennsylvania	2 July 1997 Beverly Hills, California	Gloria Hatrick	Alexander Maitland Stewart Elizabeth Ruth Jackson
Go to		James Mayflin Stewart ★★★★☆ @WeRelate.org	1907 Indiana	2 July 1977 California	Gloria	Alexander Stewart Elizabeth
Partial Matches: 6724						
Go to		Jimmy Stewart ★★★☆☆ @Geni.com	May 1908 Indiana, Pennsylvania	1997 California, Pennsylvania		
Go to		James Stewart ★★★☆☆ @MyHeritage.com	1908 Ballymoney, County Antrim, Ireland.			Robert Stewart Elizabeth Lanigan
Go to		James Stewart ★★★☆☆ @WeRelate.org	of,,Wigtownshire,Scotland			Alexander STEWART Elizabeth Douglas

Refine Search New Search Previous 1 2 3 4 5 6 7 8 9 10 Next

Figure 2: Search results as they might appear on a Linked Genealogical Data-aware website.

may be contained in multiple datasets, but because these datasets do not describe the individual in the same way and cannot link to each other, interested users will never know about the connections.

Websites that know how to utilize Linked Genealogical Data will open their users to a world of genealogical records and data previously inaccessible to them. Imagine executing a search and seeing results like those illustrated in Figure 2. The records returned could come from any site providing useful information, not just the current one. Users on different sites would be enabled to build upon and improve each other's research—a global collaboration—while still retaining ownership of their own data. A Web of Genealogical Data is the missing element needed to enable these possibilities.

8 Joining the Data Web

For a website already serving web pages related to people, places, events, etc., serving Linked Data in addition to the human-readable pages can be relatively straightforward. One technology enabling joint serving of human- and machine-readable web pages is RDFa. RDFa allows embedding of RDF statements within HTML itself.(Adida and Birbeck 2008) For example, the HTML snippet in Figure 3 both displays James Stewart's name, and encodes it as RDF. The equivalent RDF triples are also shown in Figure 3.

9 Conclusion

In this paper I have described the Semantic Web and Linked Data and advocated their use to join genealogical data together in a single, global family tree. I introduced the GEN vocabulary for expression of Linked Genealogical Data and invited feedback. I have also given an example of the impact a Web of Genealogical Data might have on user experience, and demonstrated one way of serving RDF along with HTML pages. The technology is here, the reasons for action are many, and the time for this community to connect with the Web of Data is *now!*

References

Ben Adida and Mark Birbeck. Rdfa primer: Bridging the human and data webs, 2008. URL `http://www.w3.org/TR/xhtml-rdfa-primer/`.

Tim Berners-Lee. Linked data - design issues, 2006. URL `http://www.w3.org/DesignIssues/LinkedData.html`.

Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 2011. URL `http://linkeddatabook.com/editions/1.0/`.

Graham Klyne and Jeremy J. Carroll. Resource description framework (rdf): Concepts and abstract syntax, 2004. URL

HTML:

```
<div about="http://example.com/actors/JamesStewart"
    xmlns:foaf="http://xmlns.com/foaf/0.1/"
    xmlns:gen="http://purl.org/gen/0.1#">
  <h2 about="http://example.com/names/James_Maitland_Stewart">
    <span property="foaf:givenName">James Maitland</span> 
    <span property="foaf:familyName">Stewart</span>
  </h2>
  <a rel="gen:name" resource="http://example.com/names/James_Maitland_Stewart"/>
</div>
```

RDF Equivalent:

```
<http://example.com/names/James_Maitland_Stewart> foaf:givenName
    "James Maitland" ;
  foaf:familyName "Stewart" .
<http://example.com/actors/JamesStewart> gen:name
    <http://example.com/names/James_Maitland_Stewart> .
```

Figure 3: HTML representation of James Stewart's name, with equivalent embedded, machine-readable data using RDFa.

http://www.w3.org/TR/2004/REC-rdf-concepts-20040210.

Library of Congress. Bibliographic framework initiative general plan, 2011. URL http://www.loc.gov/marc/transition/pdf/bibframework-10312011.pdf.

Frank Manola and Eric Miller. Rdf primer, 2004. URL http://www.w3.org/TR/rdf-primer.

Leo Sauermann and Richard Cyganiak. Cool uris for the semantic web, 2008. URL http://www.w3.org/TR/2007/WD-cooluris-20071217/.

Charla Woodbury and David W. Embley. Family history research on the semantic web: Building a semantic prototype for danish research. In *Family History Technology Workshop*, 2005. URL http://fht.byu.edu/prev_workshops/workshop05/FHTCD/session1/s1-CharlaWoodbury_SemanticWeb.pdf.

Ivo Zandhuis. Towards a genealogical ontology for the semantic web. URL http://www.zandhuis.nl/sw/genealogy/genont.pdf.