Using Linked Data to Create a Typological Knowledge Base

Steven Moran

Abstract In this paper, I describe the challenges in creating a Resource Description Framework (RDF) knowledge base for undertaking phonological typology. RDF is a model for data interchange that encodes representations of knowledge in a graph data structure by using sets of statements that link resource nodes via predicates that can be logically marked-up (Lassila and Swick, 1999). The model I describe uses Linked Data to combine data from disparate segment inventory databases. Once the data in these legacy databases have been made interoperable at the linguistic and computational levels, I show how additional knowledge about distinctive features is linked to the knowledge base. I call this resource the Phonetics Information Base and Lexicon (PHOIBLE)[1] and it allows users to query segment inventories from a large number of languages at both the segment and distinctive feature levels (Moran, 2012). I then show how the knowledge base is useful for investigating questions of descriptive phonological universals, e.g. "do all languages have coronals?" and "does every phonological system have at least one front vowel or the palatal glide /j/?" (Hyman, 2008).

1 Introduction

Linked Data is a set of recommended practices for publishing RDF-structured data on the Web (Bizer et al, 2007). RDF is a model for data interchange that encodes representations of knowledge in a graph data structure by using sets of statements (aka 'triples' or 'ontological commitments') that link resource nodes via predicates (Lassila and Swick, 1999; Beckett, 2004). For example the triples, graphically illustrated in Fig. 1, represent a collection of statements where a subject (the language

Steven Moran
Ludwig-Maximilians-Universität, München, Schellingstrasse 9, D-80539 München, e-mail: steve.moran@lmu.de

[1] http://phoible.org

`Sisaala, Western [ssl]`)[2] is connected via the relationship `hasSegment` to the segments `p`, `b`, `kp`, etc.

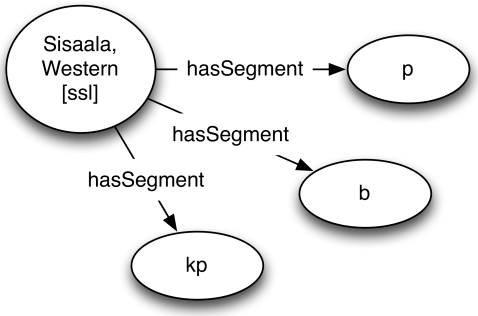

Fig. 1 Snippet of PHOIBLE RDF knowledge base

RDF is a data model for defining and specifying resources (here languages and segments) and the relations that hold between them. The product of this model, together with the technologies used to store and access it, is a knowledge base. A knowledge base encapsulates the capabilities of what several integrated technologies allow a user to achieve through the Semantic Web framework.

2 Background

Tim Berners-Lee and colleagues coined the term "Semantic Web" and gave a vision to a "web of data" (Berners-Lee et al, 2001), now commonly referred to as Linked Data. This vision of a web of interlinked Data is motivated by the fact that the Web has evolved mainly as HTML web pages that publish information for human consumption. Their inherent meaning is not interpretable by computers because they lack rich-machine readable metadata. The goal of the Semantic Web vision is to make possible the processing of information published on the Web by computers (Cardoso and Sheth, 2006). In Semantic Web architecture, an application framework stores data in a knowledge base. The Semantic Web is built in layers. A node in the Semantic Web, i.e. a concept, individual or class, is a Uniform Resource Identifier (URI). Triples are built with URIs that define the subject, predicate and object of a statement. Each triple describes a fact. The subject and predicate are defined with a URI. The object of the statement can be either a URI or some other definable data type, such as a string literal or an integer. The URI is a key feature in the overall architecture because each provides a unique identifier within a global namespace. Since triples are built with URIs, they can be easily merged from many different sources via common URIs or defining of relationships between URIs via additional

[2] ISO 639-3 language codes are given in brackets [ ].

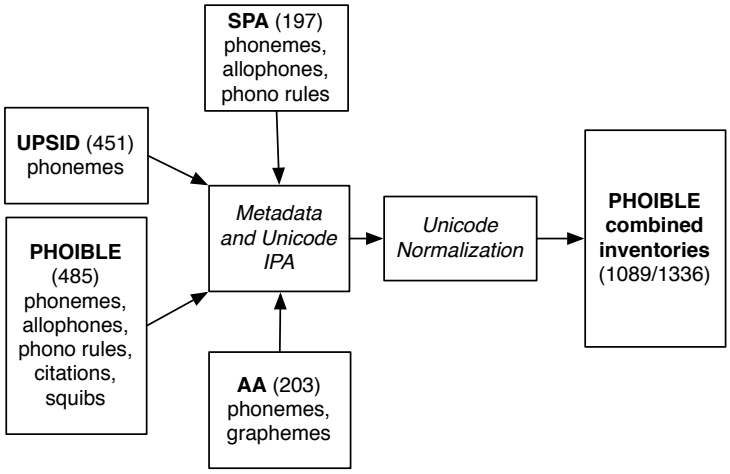

Fig. 2 Resources combined into PHOIBLE

triples. The URI is a key design feature because it provides the mechanism for global naming and connects each resource in the statement to a Web resource.

3 Data and Model

Segment inventory databases have been used since the 1970s to investigate phonological universals and to report on the distribution of sounds in the world's languages. The data used in this paper come from three databases and from my own work extracting segment inventories from the linguistics literature. The first database is the Stanford Phonology Archive (SPA; Crothers et al, 1979). It contains segment inventories (phonemes and allophones) for 197 languages. The second is the UCLA Phonological Segment Inventory Database (UPSID), which contains 451 languages' phoneme inventories (Maddieson, 1984; Maddieson and Precoda, 1990). The third is the 200 language sample (phonemes and graphemes) from *Alphabets des langues africaines* (Hartell, 1993; Chanard, 2006).[3] Additional inventories, extracted from over 400 grammars or phonological descriptions, include phonemes, allophones and phonological rules. Figure 2 illustrates how these resources are combined into the PHOIBLE knowledge base. At the time of writing, there are 1336 segment inventories, which represent 1089 distinct languages.

Combining disparate data sets is challenging. Integrating segment inventories from different resources into one interoperable data set poses three main challenges.

[3] Note that the quality of the original and digitized versions of these data were questionable, so Christopher Green and I corrected and/or collected additional resources for these inventories.

The first challenge is how to compare segments from different transcription systems. This issue is addressed by interpreting transcription systems' segments into the International Phonetic Alphabet (IPA; International Phonetic Association, 2005), which I use as an interlingual pivot. However, re-encoding segments into IPA can also be problematic. For example, when more than one diacritic appears to the right of, or below a segment, in which order should they appear (e.g. a creaky voiced syllabic dental nasal / n̪̰̬ /)? A 'correct' diacritic ordering does not seem to be explicitly stated in the IPA. Therefore, I chose to create an order that all (applicable) segments in all segment inventory databases in PHOIBLE abide by. For example, when there is more than one diacritic to the right of a segment, the order is: unreleased/lateral release/nasal release → palatalized → labialized → velarized → pharyngealized → aspirated/ejective → long, e.g. a labialized aspirated long alveolar plosive is represented as / t^{wh}ː /.[4]

The second challenge involves making segments interoperable at the computational level. For example, a nasalized creaky vowel /ḛ̃/ has diacritics that appear above and below the vowel. Although the order is not visually distinguishable, computationally there are two sequences in which these characters can occur (depending on how they are keyboarded by the linguist).[5] To address this issue, Unicode normalization is needed to decompose each complex segment into an ordered sequence of characters to ensure that equivalent strings have the same binary representation (The Unicode Consortium, 2007).

The third challenge involves associating metadata with each segment inventory in the knowledge base. Each inventory needs to be identified with a language and bibliographic information about the publication from which the inventory was gathered. In my approach, each inventory is identified by an ISO 639-3 language code, which allows inventories representing the same language to be compared systematically, even if the language names used in those resources differ (e.g. *German* vs *Deutsch*). Bibliographic data is associated with the Open Language Archives Community (OLAC) metadata set. OLAC expands the set of DCMI metadata categories to include information pertinent to linguistics data to create a standard way to document many types of language resources, by adding metadata elements like subject language and linguistic data type to enhance greater discovery of language resources.

After these challenges have been addressed to make the data interoperable at the linguistic and computational levels, RDF is used to model languages and their segment inventories. The RDF data structure represents a collection of facts about information using URIs.[6] This knowledge base can be queried with SPARQL, an RDF

[4] Ordering conventions are stated explicitly in Moran 2012, as are the SPA and UPSID segment-to-IPA mappings.

[5] There are two possibilities in Unicode: U+0065 + U+0330 + U+0303 (LATIN SMALL LETTER E + COMBINING TILDE BELOW + COMBINING TILDE) or U+0065 + U+0303 + U+0330 (LATIN SMALL LETTER A + COMBINING TILDE + COMBINING TILDE BELOW).

[6] The figures in this paper use terms to represent URIs for readability purposes. However, in the published PHOIBLE RDF model, a segment is defined by an explicit URI, for example `http://phoible.org/segment/kp`, which is consistent with the RDF specification.

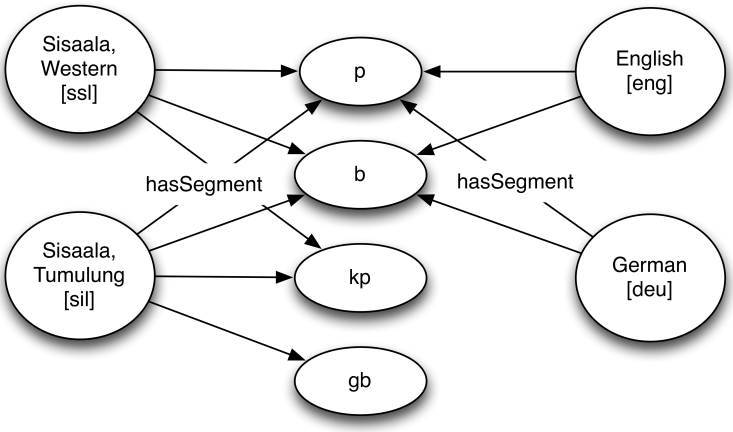

Fig. 3 Expanded snippet of PHOIBLE RDF knowledge base

query language (Prud'Hommeaux and Seaborne, 2006). SPARQL queries consist of triple patterns that match concepts and their relations by binding variables to match graph patterns. For example, a SPARQL query on the knowledge base in Fig. 3 to retrieve the segments of `Sisaala, Western [ssl]` is given in example 2:

(2)
```
SELECT ?segments
WHERE {ssl hasSegment ?segments}
```

The SPARQL query matches any sets of triples that contain `ssl` (for Sisaala, Western) as the subject and `hasSegment` as the predicate. In this snippet the query would return the segments `p`, `b` and `kp`.

An important feature of the RDF graph model, and thus Linked Data, is that the the same knowledge representation language is used in the knowledge base's structure and its encoding of data instances. This is because the knowledge base uses triples to define its structure. Thus, triples can be easily added to by defining new resources (subjects or objects of triples) or predicates. This self-describing structure supports a model of open and shared data. For example, if one wants to add new knowledge about distinctive features to the segment inventory knowledge base, the features can be added to the graph by linking them from each segment via another predicate.[7] For example, I define a URI for the `hasFeature` predicate to link features to segments in Fig. 4. Now if the user wants to query for segments in languages that only contain certain features, they can. Perhaps someone wants to query for all languages that have velar plosive segments. Then they could use the query given in

[7] Distinctive feature sets lack typological coverage for the vast variety of segment types that appear in the linguistics literature. In this paper I gloss over the challenges involved in assigning features to segments. See Moran 2012 for details.

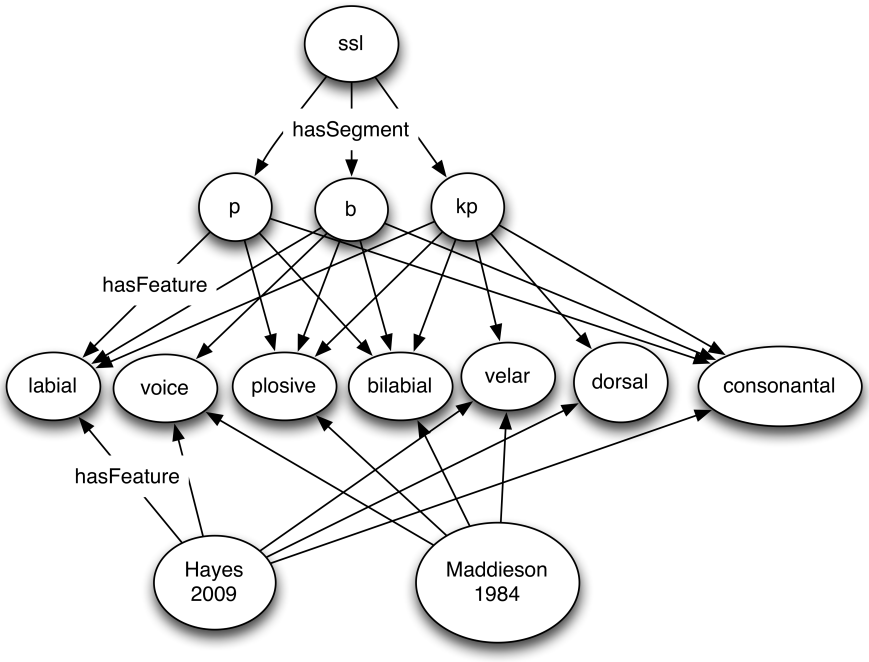

Fig. 4 Adding distinctive features to the knowledge base

example 3. It is now apparent how one might use the PHOIBLE Linked Data graph to investigate questions of phonological universals.

(3)
```
SELECT ?languages
WHERE {?languages hasSegment ?segments .
?segments hasFeature plosive .
?segments hasFeature velar }
```

4 Query

The PHOIBLE Linked Data graph of segments and distinctive features can be used to investigate descriptive universals of phonological inventories, for example, some of those stated in Hyman 2008.[8] One area to investigate is vowel systems. Hyman asks if "every phonological system has at least one unrounded vowel" and reaches

[8] See Hyman 2008 for a description of various types of universals. At this time, PHOIBLE cannot be used to address theory-dependent architectural universals, e.g. statements made within Optimality Theory, or universals dealing with tendencies above the segment level, e.g. universals regarding syllable structure.

the conclusion, based on the data from UPSID-451 (Maddieson, 1984; Maddieson and Precoda, 1990),[9] that no language in UPSID-451 has less than two unrounded vowels. This universal can be formulated, as in the SPARQL query in example 4, using features from Hayes (2009).

(4)
```
SELECT DISTINCT ?languages
WHERE {?languages phoible:hasSegment ?segments .
?segments phoible:hasFeature feature:SYLLABIC .
?segments phoible:notHasFeature feature:CONSONANTAL .
?segments phoible:notHasFeature feature:ROUND }
```

This query selects all distinct languages that have segments that have the features [+SYLLABIC], [−CONSONANTAL] and [−ROUND], i.e. unrounded vowels. All 1089 languages are returned. The query can also be modified to return all languages and their segments, shown in example 5.

(5)
```
SELECT ?languages ?segments
WHERE {?languages phoible:hasSegment ?segments .
?segments phoible:hasFeature feature:SYLLABIC .
?segments phoible:notHasFeature feature:CONSONANTAL .
?segments phoible:notHasFeature feature:ROUND }
```

This query returns all language inventories and their unrounded vowels according to our feature set and specification.

Next, by querying the graph for distinct languages that have segments that have the features [+SYLLABIC], [−CONSONANTAL] and [+BACK], we can confirm that Hyman's stated universal, "every phonological system has at least one back vowel", holds in the expanded PHOIBLE dataset.

(6)
```
SELECT DISTINCT ?languages
WHERE {?languages phoible:hasSegment ?segments .
?segments phoible:hasFeature feature:SYLLABIC .
?segments phoible:notHasFeature feature:CONSONANTAL .
?segments phoible:hasFeature feature:BACK}
```

Another possible universal investigated by Hyman is "every phonological system has at least one front vowel or the palatal glide /j/". This can be asked of the PHOIBLE Linked Data graph by using the SPARQL UNION operator to query all languages that have segments of a particular feature make-up ([+SYLLABIC, +ROUND, −CONSONANTAL] or the segment /j/.[10] This universal also holds in the PHOIBLE dataset.

(7)
```
SELECT DISTINCT ?languages
WHERE {?languages phoible:hasSegment ?segments .
?segments phoible:hasFeature feature:SYLLABIC .
?segments phoible:hasFeature feature:FRONT .
?segments phoible:notHasFeature feature:CONSONANTAL .
UNION {?languages phoible:hasSegment segment:j}}
```

[9] The data is taken from Henning Reetz's online version, at: http://web.phonetik.uni-frankfurt.de/UPSID.html.

[10] Note I use IPA /j/ instead of /y/.

Another area to investigate descriptive universals in segment inventories is in consonant systems. Hyman posits that "every phonological system has stops" and "every phonological system has coronal phonemes". Example 8 queries for the first universal by selecting all languages that have segments that have the feature [−DELAYED_RELEASE], i.e. all stops. Indeed all languages in the PHOIBLE dataset have at least one stop.

(8)
```
SELECT DISTINCT ?languages
WHERE {?languages phoible:hasSegment ?segments .
?segments phoible:notHasFeature feature:DELAYED\_RELEASE }
```

Finally, the query in 9 checks the PHOIBLE data set for any languages that do not have a coronal phoneme.

(9)
```
SELECT DISTINCT ?languages
WHERE {?languages phoible:hasSegment ?segments .
?segments phoible:hasFeature feature:CORONAL }
```

The knowledge base, however, contains counter-evidence to the universal, found in the segment inventory of Northwest Mekeo [mek] Jones (1995, 1998), which has the consonants: / p, β, m, w, ɡ, ŋ, j /. Northwest Mekeo's lack of coronals was reported in Blevins 2009, which was inspiration to compile a larger set of segment inventories than UPSID-451 and to develop a typological knowledge base using Linked Data; thus allowing users to query at the levels of segments and distinctive features to investigate questions regarding phonological typology and universals.

5 Conclusion

The knowledge base is a data-centric model. In comparison to individually devised relational databases, the knowledge base facilitates data sharing by publishing a self-describing data model according to explicitly encoded relationships found in the data. This graph data model is more dynamic and allows information to be added at any node. In my opinion, the graph data structure uses a technology that embraces principles towards a cyberinfrastructure approach (cf. Bender and Langendoen, 2010; Pericliev, 2010). The major benefit of using Linked Data is that the graph data structure is designed explicitly for data sharing. Because of global scope, the triple structure that makes up the graph allows for easy information integration. Two graphs from different sources that share a given URI can be merged without transforming the data. Another benefit, not explored in this paper, is the ability to mark-up RDF predicates with logical constructions using the Web Ontology Language (OWL; McGuinness and van Harmelen, 2004). For example, if a user does not consider length a phonemic property of segment inventories, then he or she can use the OWL property 'owl:sameAs' to state that the feature [+LONG] is equivalent to [−LONG]. With one simple statement the user can change the contents of the knowledge base, without changing the underlying data in the Linked Data graph.

Acknowledgements Thanks to Dan McCloy and Richard Wright for assisting me in mapping the SPA and UPSID segments into IPA and to Christopher Green for help with the collection and analysis of segment inventories from African languages. Thanks also to to three anonymous reviewers for feedback.

References

Beckett D (2004) RDF/XML Syntax Specification (Revised). Tech. rep., W3C, URL http://www.w3.org/TR/rdf-syntax-grammar/

Bender EM, Langendoen DT (2010) Computational Linguistics in Support of Linguistic Theory. Linguistic Issues in Language Technology (LiLT) 3(2):1–31

Berners-Lee T, Hendler J, Lassila O (2001) The Semantic Web. Scientific American URL http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

Bizer C, Cyganiak R, Heath T (2007) How to publish linked data on the web. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

Blevins J (2009) Another Universal Bites the Dust: Northwest Mekeo Lacks Coronal Phonemes. Oceanic Linguistics 48(1):264–273

Cardoso J, Sheth AP (eds) (2006) Semantic Web Services, Processes and Applications. Springer

Chanard C (2006) Systèmes alphabétiques des langues africaines. http://sumale.vjf.cnrs.fr/phono/

Crothers JH, Lorentz JP, Sherman DA, Vihman MM (1979) Handbook of phonological data from a sample of the world's languages: A report of the stanford phonology archive

Hartell RL (ed) (1993) Alphabets des langues africaines. UNESCO and Société Internationale de Linguistique

Hayes B (2009) Introductory Phonology. Blackwell

Hyman LM (2008) Universals in phonology. The Linguistic Review 25:83–137

International Phonetic Association (2005) International Phonetic Alphabet. Tech. rep., International Phonetic Association, URL http://www.arts.gla.ac.uk/IPA/

Jones AA (1995) Mekeo. In: Tryon DT (ed) Comparative Austronesian Dictionary: An Introduction to Austronesian Studies, Part 1: Fascicle 2, Mouton de Gruyter

Jones AA (1998) Towards a Lexicogrammar of Mekeo (An Austronesian Language of Western Central Papua). Pacific Linguistics, Canberra

Lassila O, Swick RR (1999) Resource Description Framework (RDF): Model and syntax specification (recommendation). http://www.w3.org/TR/REC-rdf-syntax

Maddieson I (1984) Pattern of Sounds. Cambridge University Press, Cambridge, UK

Maddieson I, Precoda K (1990) Updating UPSID. In: UCLA Working Papers in Phonetics, vol 74, pp 104–111

McGuinness DL, van Harmelen F (2004) OWL Web Ontology Language Overview. URL http://www.w3.org/TR/owl-features/

Moran S (2012) Phonetics information base. PhD thesis, University of Washington
Pericliev V (2010) Machine-Aided Linguistic Discovery: An Introduction and Some Examples. London: Equinox
Prud'Hommeaux E, Seaborne A (2006) SPARQL query language for RDF. W3C working draft 4
The Unicode Consortium (2007) The Unicode Standard, Version 5.0.0, defined by: The Unicode Standard, Version 5.0. URL http://www.unicode.org/versions/Unicode5.0.0/