Undefined 1 (2009) 1–5
IOS Press

Linguistic Resources Enhanced with Geospatial Information

Richard Littauer [a,b], Boris Villazon-Terrazas [c] and Steven Moran [d,e]

[a] *Department of Intelligent Computer Systems, University of Malta, Msida, MSD2080, Malta*
[b] *Computational Linguistics Department, Saarland University, Saarbrücken, 66121, Germany*
E-mail: littauer@coli.uni-saarland.de
[c] *Intelligent Software Components, iSOCO, S.A., Av. del Partenon 16-18, Madrid, Spain*
E-mail: bvillazon@isoco.com
[d] *Department of Linguistics, University of Zürich, Plattenstrasse 54, CH-8032 Zürich, Switzerland*
[e] *Research Unit Quantitative Language Comparison, Ludwig Maximilian University, Geschwister Scholl Platz 1, D-80539 Munich, Germany*
E-mail: steve.moran@lmu.de

Abstract.
In this short report on language data and RDF tools, we describe the transformation process that we undertook to convert spreadsheet data about a group of endangered languages and where they are spoken in West Africa into an RDF triple store. We use RDF tools to organize and visualize these data on a world map, accessible through a web browser. The functionality we develop allows researchers to see where these languages are spoken and to query the language data. This type of development not only showcases the power of RDF, but it provides a powerful tool for linguists trying the solve the mysteries of the genealogical relatedness of the Dogon languages.

Keywords: Semantic Web, Linked Data, LLOD, Linguistics, Typology, Language Resources, Geospatial Mapping

1. Introduction

Linked Data presents many opportunities to access and share data in different formats and for different purposes. In linguistics and related fields like cultural archaeology and population genetics, visualization of data points on maps is particularly beneficial in formulating hypotheses about data sets, particularly sparse ones, which is often the case in these fields. In this short report, we describe how we converted a spreadsheet that contains information about endangered Dogon languages and where they are spoken in small rural villages in Mali, West Africa, into an Resource Description Framework (RDF) triple store so that we could leverage other RDF tools to visualize these data. The result gives researchers a clearer picture of the dispersal of Dogon speakers and we show that the spreadsheet-to-RDF conversion pipeline that we develop is applicable to any data set that can be combined with GPS coordinates.

2. Background

In the visualization of language data, there has been work on displaying language differences on a broad scale, including presenting hierarchical and cross-linguistic data [7,8], displaying related languages gathered from the World Atlas of Language Structures (WALS) by geographical proximity and relatedness [6], displaying word meanings on a map [9], and displaying the location of languages that contain some type of typological feature language locations on a world map [5]. There has also recently been visualizations that display language relatedness and dialectology using lexical items and location together [11].

0000-0000/09/$00.00 © 2009 – IOS Press and the authors. All rights reserved

In this work we derive RDF from simple table data stored in a spreadsheet, leverage the ability of RDF graphs to be easily merged, and harness different RDF tools to display geospatial data in the map4rdf software, which is freely available and runs in the browser. In doing so, we provide detailed information about the location of villages in Mali in which Dogon languages are spoken. Dogon is an interesting language family because until recently there was very little that was known about these languages. In fact, until as late as 1989, Dogon appeared in reference books on African languages as if it were a single language (cf. [1,2]). Current estimates from experts working in Mali is that there are now over 20 mutually unintelligble Dogon languages, with new varieties being "discovered" every year. However, the current genealogical relatedness of these languages is still unclear, as is the internal structure of the Dogon language family. Additionally, due to the typological characteristics of Dogon languages, such as these languages' lack of noun classes that are typical of sub-saharan West African languages in general or Dogon's SOV basic word order (instead of SVO like many of its neighbors), the position of the Dogon language family relative to other African language families remains unclear. Thus in disentangling the mysteries of how Dogon languages are related within their family, an interactive visual reference of where the languages are spoken is a useful tool for exploring avenues of possible genealogical decent due to geographic proximity and other effects like borrowing due to areal contact.

3. LLD Life Cycle

In this section we present the specification of the Linked Data Life Cycle presented in [10] as applied to linguistic resources to visualize them with geospatial information.

3.1. Linguistics Resources

Our data source consists of a spreadsheet containing GPS coordinates of villages where the different Dogon languages are spoken in Mali. It also contains information about each of these languages, such as the language name, ISO 639-3 language name identifier, the language family and family code, village name, etc. and it can be easily combined via ISO 639-3 codes with dictionary data from each language. These datasets come from the Dogon Languages Project and are freely available online.[1] Each set of data points per village is associated with a GPS coordinate and can thus be plotted on a world map. Because the set of Dogon languages that belong to the Dogon language family have been until recently poorly documented and described, information about where these languages are spoken in relation to each other can assist linguists in identifying the genealogical relatedness of these languages. The visualization of linguistic information on maps has been a successful method for generating and testing hypotheses (cf. [5]).

3.2. Specification

The process of publishing Linked Data has an iterative incremental life cycle model. Data sources must be identified and analyzed and entities in the data must be assigned a URI. A key element of Linked Data is also the ability to reuse and leverage data that has already been published as Linked Data. By identifying the schema of resources that are to be transformed into Linked Data, conceptual components and their relationships can be properly modeled into the RDF triple format. In the Dogon GPS spreadsheet, we were able to identify fields such as language name, ISO 639-3 code, language family and subfamily, alternative languages spoken in each village, village names, municipality, notes about the speaker's society, and geospatial information and assign them a URI. See Fig. 1.

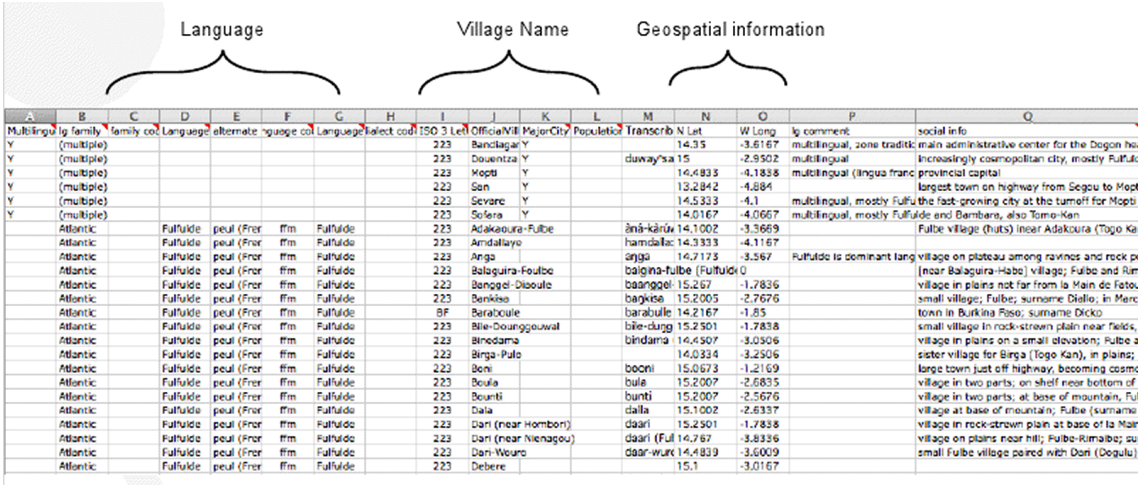
All resources in the dataset are given dereferenceable URIs and we've attempted to use meaningful names instead of opaque ones. We also reuse URIs where we can, including using the General Ontology of Linguistic Description (GOLD) for morphosyntactic data descriptions [4].[2] The base URI structure uses the `http://linguistic.linkeddata.es/` namespace. Vocabulary elements are appended with `/ontology/{property|class}` and instances with `/dataset/resource/{r.type|r.name}`. We also reused URIs from the WGS84 Geo Basic Vocabulary for the representation of geospatial data.[3]

3.3. RDF Generation

Next, the spreadsheet data was transformed into RDF. First we imported the spreadsheet into MySQL. Then we defined a set of R2RML mappings. R2RML

[1] `http://dogonlanguages.org`
[2] `http://linguistics-ontology.org/`.
[3] `http://www.w3.org/2003/01/geo/`

Fig. 1. Data that contains villages in Mali with language information



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
Multilingu	lg family	family co	Language	alternate	nguage co	Language	dialect cod	ISO 3 Let	OfficialVil	MajorCity	Populatio	Transcrib	N Lat	W Long	lg comment	social info
Y	(multiple)							223	Bandiagar	Y			14.35	-3.6167	multilingual, zone traditic	main administrative center for the Dogon he
Y	(multiple)							223	Douentza	Y		duway'sa	15	-2.9502	multilingual	increasingly cosmopolitan city, mostly Fulfuld
Y	(multiple)							223	Mopti	Y			14.4833	-4.1838	multilingual (lingua franc	provincial capital
Y	(multiple)							223	San	Y			13.2842	-4.884		largest town on highway from Segou to Mopt
Y	(multiple)							223	Sevare	Y			14.5333	-4.1	multilingual, mostly Fulfu	the fast-growing city at the turnoff for Mopti
Y	(multiple)							223	Sofara	Y			14.0167	-4.0667	multilingual, mostly Fulfulde and Bambara, also Tomo-Kan	
	Atlantic		Fulfulde	peul (Fren	ffm	Fulfulde		223	Adakaoura-Fulbe			ànà-kàrúv	14.1002	-3.3669		Fulbe village (huts) linear Adakoura (Togo Ka
	Atlantic		Fulfulde	peul (Fren	ffm	Fulfulde		223	Amdallaye			hamdallay	14.3333	-4.1167		
	Atlantic		Fulfulde	peul (Fren	ffm	Fulfulde		223	Anga			anga	14.7173	-3.567	Fulfulde is dominant lang	village on plateau among ravines and rock pi
	Atlantic		Fulfulde	peul (Fren	ffm	Fulfulde		223	Balaguira-Foulbe			baigina-fulbe (Fulfulde	0			[near Balaguira-Habe] village; Fulbe and Rim
	Atlantic		Fulfulde	peul (Fren	ffm	Fulfulde		223	Banggel-Dioulé			baanggel	15.267	-1.7836		village in plains not far from la Main de Fatou
	Atlantic		Fulfulde	peul (Fren	ffm	Fulfulde		223	Benkiso			bangkisa	15.2005	-2.7676		small village; Fulbe; surname Diallo; in Merc
	Atlantic		Fulfulde	peul (Fren	ffm	Fulfulde		BF	Baraboule			barabulle	14.2167	-1.85		town in Burkina Faso; surname Dicko
	Atlantic		Fulfulde	peul (Fren	ffm	Fulfulde		223	Bile-Dounggouwal			bile-dugg	15.2501	-1.7838		small village in rock-strewn plain near fields,
	Atlantic		Fulfulde	peul (Fren	ffm	Fulfulde		223	Binedama			bindama	14.4507	-3.0506		village in plains on a small elevation; Fulbe a
	Atlantic		Fulfulde	peul (Fren	ffm	Fulfulde		223	Birga-Pulo				14.0334	-3.2506		sister village for Birga (Togo Kan), in plains;
	Atlantic		Fulfulde	peul (Fren	ffm	Fulfulde		223	Boni			booni	15.0673	-1.2169		large town just off highway, becoming cosmo
	Atlantic		Fulfulde	peul (Fren	ffm	Fulfulde		223	Boula			bula	15.2007	-2.6835		village in two parts; on shelf near bottom of
	Atlantic		Fulfulde	peul (Fren	ffm	Fulfulde		223	Bounti			bunti	15.2007	-2.5676		village in two parts; at base of mountain, Ful
	Atlantic		Fulfulde	peul (Fren	ffm	Fulfulde		223	Dala			dalla	15.1002	-2.6337		village at base of mountain; Fulbe (surname
	Atlantic		Fulfulde	peul (Fren	ffm	Fulfulde		223	Dari (near Hombori)			daari	15.2501	-1.7838		village in rock-strewn plain at base of la Main
	Atlantic		Fulfulde	peul (Fren	ffm	Fulfulde		223	Dari (near Nienagou)			daari (Ful	14.767	-3.8336		village on plains near hill; Fulbe-Rimaibe; su
	Atlantic		Fulfulde	peul (Fren	ffm	Fulfulde		223	Dari-Wouro			daar-wur	14.4839	-3.6009		small Fulbe village paired with Dari (Dogulu)
	Atlantic		Fulfulde	peul (Fren	ffm	Fulfulde		223	Debere				15.1	-3.0167		

is a RDF-to-RDF mapping language and we used it to creat mappings between elements in the MySQL database table from the spreadsheet and the RDF vocabulary that we defined.[4] Lastly, using the R2RML engine and morph,[5] we generated the RDF instances using the R2RML defined mappings.

3.4. Publication

The RDF data that we generated is stored in a triple store with the Virtuoso software, which we use to publish the data online.[6] Integrated with Pubby,[7] Virtuoso allows us to leverage content management to serve up machine-readable and human consumable webpages that contain information about each village, such as which languages are spoken there, where the village is located, additional information about the society, etc.[8] Virtuoso also provides a SPARQL endpoint with which we can query and share the data.

3.5. Exploitation

Following the previous steps of specification, RDF generation and publication, we expose the RDF data, enhanced with GPS coordinates, using map4rdf.[9] map4rdf is a maps viewer of RDF resources with geometrical information built on OpenStreetMap[10] and it can be used to visualize information in RDF datasets. Additionally, it is extensible with Google app plugins. The parameters of map4rdf must be set so that the application knows where to locate the endpoint of Dogon data in RDF (that we set up with Virutoso) and which geometry model that we are using (since there different standards for geo-mapping). With the parameters set, a user can open the map4rdf application in his or her web browser and explore the location of villages where Dogon are spoken.[11] Fig. 2 provides an illustration.

Each point on the map comes from GPS coordinates in the original spreadsheet, which have been transformed into RDF triples and stored in a triple store with Virtuoso. This triple store can be queried with SPARQL or its endpoint can be given as an endpoint for programs like map4rdf to access its data contents. Each pin in Fig. 2 can be clicked on, showing the village name, its latitude and longitude, and a link for more information about the language. This is shown in Fig. 3.

When clicking on the link for more information, a request is sent to the SPARQL endpoint for all information in the RDF triple store about that particular village. When accessing the data through map4rdf, the endpoint knows through content management to re-

[4] http://www.w3.org/TR/r2rml/

[5] https://github.com/boricles/morph

[6] http://virtuoso.openlinksw.com/

[7] http://www4.wiwiss.fu-berlin.de/pubby/

[8] See for example the page on the village Boni: http://linguistic.linkeddata.es/page/mlode/resource/Village/Boni.

[9] https://github.com/boricles/linked-data-visualization-tools

[10] http://www.openstreetmap.org/

[11] The map4rdf instantiation for the Dogon villages resides at: http://geo.linkeddata.es/map4rdf-dogon/.

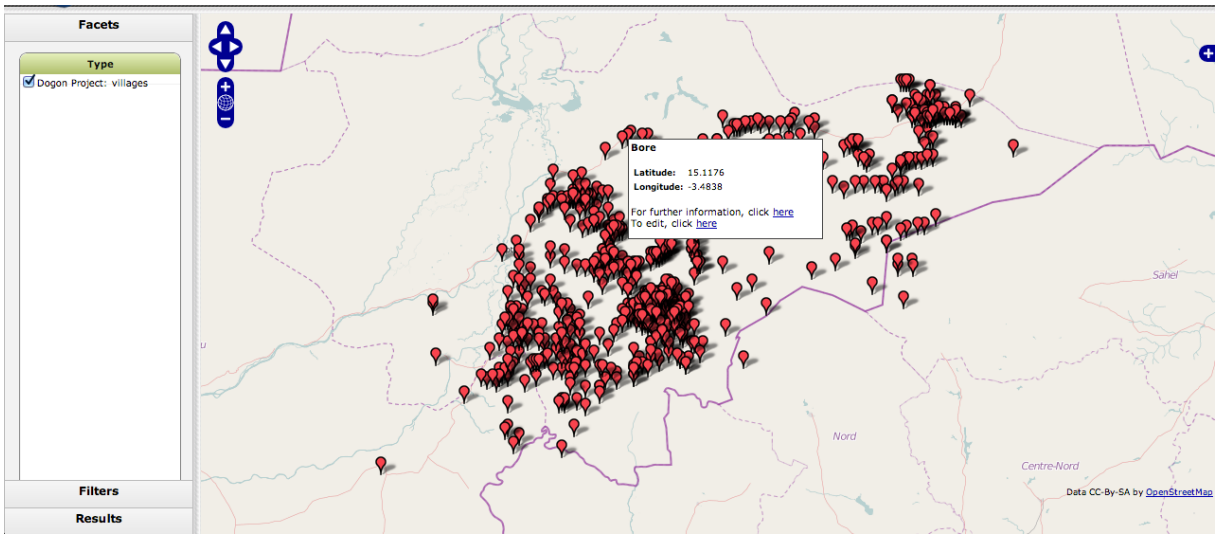


Fig. 2. Visualization of the Dogon villages

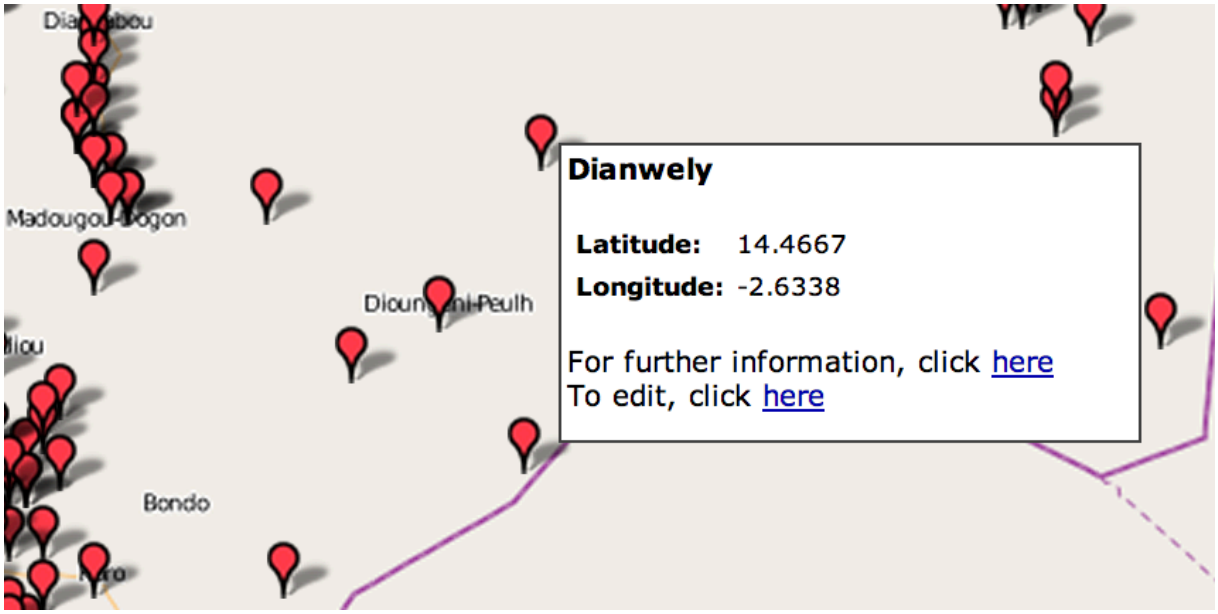
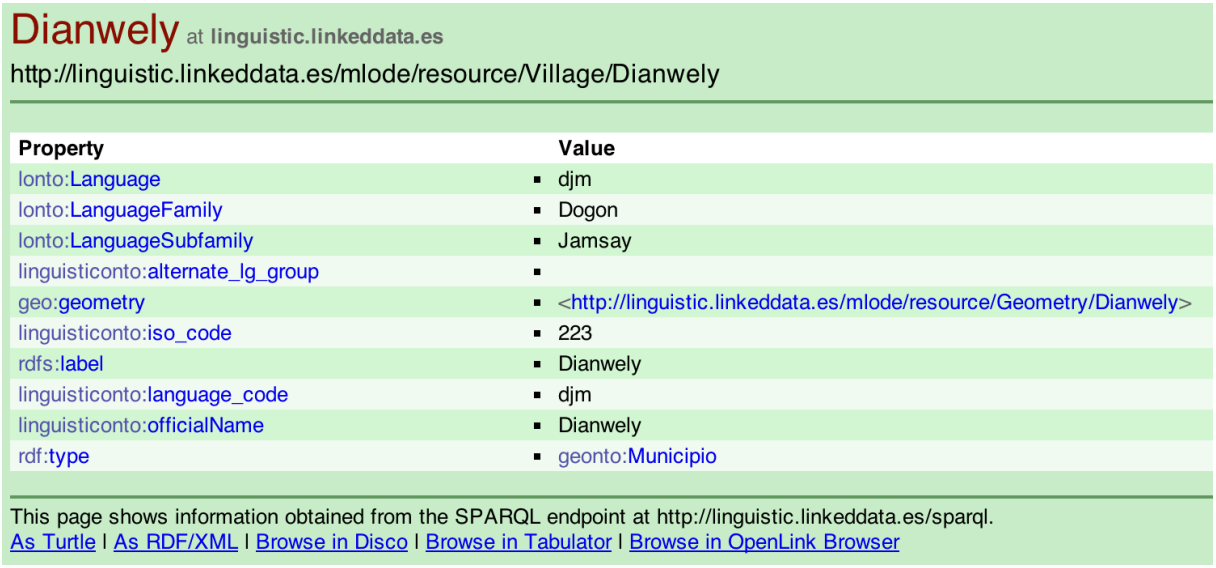


Fig. 3. Clicking on a pin

turn an HTML page that displays the query results, as shown in Fig. 4.

4. Summary

We have briefly shown here a workflow to transform data from a simple spreadsheet into an RDF triple store that can queried using a SPARQL endpoint, and an application called map4rdf that uses this endpoint with GPS coordinates to visualize RDF data on a world map. Moreover, the tools that we have used here are open source and freely available. Converting linguistic data into RDF can be a straightforward process and we have shown the steps and some tools to assist in that transformation. There is much data available about languages and their typological features on the Web, which are often available in simple .csv formats. For example, the contents of World At-



Dianwely at **linguistic.linkeddata.es**

http://linguistic.linkeddata.es/mlode/resource/Village/Dianwely

Property	Value
lonto:Language	djm
lonto:LanguageFamily	Dogon
lonto:LanguageSubfamily	Jamsay
linguisticonto:alternate_lg_group	
geo:geometry	<http://linguistic.linkeddata.es/mlode/resource/Geometry/Dianwely>
linguisticonto:iso_code	223
rdfs:label	Dianwely
linguisticonto:language_code	djm
linguisticonto:officialName	Dianwely
rdf:type	geonto:Municipio

This page shows information obtained from the SPARQL endpoint at http://linguistic.linkeddata.es/sparql.
As Turtle | As RDF/XML | Browse in Disco | Browse in Tabulator | Browse in OpenLink Browser

Fig. 4. More information about a village

las of Language Structures (WALS)[12] [5] have been converted from .csv to RDF and are available through the MLODE SPARQL endpoint.[13] It was a trivial task for us to set up map4rdf to point at the WALS RDF data, so that we could also visualize its contents, which contain over 2000 languages' data points. Whereas the online version of WALS already contains maps of typological features of languages, their use is limited and by leveraging RDF as we have with WALS and the Dogon data, we can easily combine these disparate datasets, so that, for example, we can merge data about languages and their typological features from both datasets. This allows us to visualize not only the villages where Dogon languages are spoken, but linguistic features of languages spoken in this area of Mali encoded in WALS. This mashup provides even more detailed information about the features of these different languages, which provides another important data source in untangling the mystery of why Dogon languages are so different than other language families in West Africa. It also shows the power of encoding data in RDF and leveraging RDF tools.

References

[1] J. Bendor-Samuel, E. J. Olsen, and A. R. White. Dogon. In J. Bendor-Samuel, editor, *The Niger-Congo Languages–A Classification and Description of Africa's Largest Language Family*, pages 169–177. University Press of America, Lanham, Maryland, 1989.

[2] R. Blench. A survey of Dogon languages in Mali: overview. *OGMIOS*, 26:14–15, 2005.

[3] A. de León, F. Wisniewki, B. Villazón-Terrazas, and O. Corcho. Map4rdf - Faceted Browser for Geospatial Datasets. In *Proceedings of the First Workshop on USING OPEN DATA*. W3C, June 2012.

[4] S. Farrar and D. T. Langendoen. A Linguistic Ontology for the Semantic Web. *GLOT International*, 7:97–100, 2003.

[5] M. Haspelmath, M. Dryer, D. Gil, and B. Comrie, editors. *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich, 2008.

[6] R. Littauer, R. Turnbull, and A. Palmer. Visualising typological relationships: Plotting wals with heat maps. In *Proceedings of the EACL 2012 Workshop on the Visualization of Linguistic Patterns*, page 4, Avignon, France, April 2012. Association for Computational Linguistics.

[7] C. Rohrdantz, M. Hund, T. Mayer, B. Wälchli, and D. A. Keim. The world's languages explorer: Visual analysis of language features in genealogical and areal contexts. *Comp. Graph. Forum*, 31(3pt1):935–944, June 2012.

[8] C. Rohrdantz, T. Mayer, M. Butt, F. Plank, and D. A. Keim. Comparative visual analysis of cross-linguistic featuretures. In J. Kohlhammer and D. A. Keim, editors, *Proceedings of the International Symposium on Visual Analytics Science and Technology (EuroVAST 2010). The DEFINITIVE VERSION is available at diglib.eg.org.*, pages 27–32, 2010.

[9] R. Therón, L. Fontanillo, A. Esteban, and C. Segun. Visual analytics: A novel approach in corpus linguistics and the nuevo diccionario histórico del español. In *III Congreso Internacional de Lingstica de Corpusi*, 2011.

[10] B. Villazón-Terrazas, L. Vilches-Blázquez, O. Corcho, and A. Gómez-Pérez. Methodological Guidelines for Publishing Government Linked Data Linking Government Data. In

[12] http://wals.info
[13] http://mlode-sparql.nlp2rdf.org/sparql

D. Wood, editor, *Linking Government Data*, chapter 2, pages 27–49. Springer New York, New York, NY, 2011.

[11] M. Wieling, J. Nerbonne, and R. H. Baayen. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE*, 6(9):e23613, 09 2011.