Toward a Lexicographic Infrastructure for Serbian Digital Libraries

Natalia Ermolaev, Ph.D.
M.L.I.S. Candidate, Rutgers School of Communication and Information
Rutgers, The State University of New Jersey
4 Huntington St.
New Brunswick, NJ 08901-1071 USA
Email: nata3@eden.rutgers.edu

Toma Tasovac, M.A.
Director, Belgrade Center for Digital Humanities
Ph.D. Candidate, Trinity College Dublin, Digital Arts and Humanities
Address: Džordža Vašingtona 28A
11000 Beograd, Serbia
Email: ttasovac@humanistika.org

Digital infrastructures are essential for advancing humanities scholarship in the digital age. In this paper, we discuss three ongoing projects by the Belgrade Center for Digital Humanities (BCDH) — “Transpoetika,” the “Digital Library of Serbian Cultural Heritage of the 18th Century” and the “Annotated Digital Edition of Vuk Stefanović Karadžić’s Srpski rječnik” — as case studies for how a lexicographic infrastructure can contribute to the creation of more sophisticated digital libraries (DLs). Designed around a robust and interoperable dictionary web service, the BCDH DLs offer a broad range of functionalities for user engagement with cultural heritage texts. Features such as full-content searching, multiple browsing options, glosses, concordances and indexes, hypertext and hypermedia links allow users to define, explain, and order the language and literary culture. At the same time, these DLs highlight the inherent contingency and dialogic nature of language systems by inviting users to comment, question, explore, and play with the material through features such as innovative interface and social media. By facilitating exploration of the complexities and variability of cultural heritage, the BCDH projects exemplify the positive potentialities of DLs for humanities scholarship and cultural heritage discourse.

Introduction

While the first wave of digital libraries (DLs) was characterized by the creation of static, content-centric systems of storage and retrieval of centrally located texts, the DLs of the future promise to facilitate information exchange, collaboration and the synthesis of various media to inspire new ways of “defining, managing, and interacting with the cultural and intellectual record of our societies” (DELOS, 2007; Lynch, 2011).

As today’s DLs move toward realizing this “grander vision” (Greenstein and Thorin, 2002), their creators face a series of crucial technological, intellectual and cultural questions about how DLs will shape scholarly pursuits and cultural heritage discourse (Borgman, 2009; Dalbello, 2009). With an inherent focus on the diversity of the human record and an approach to the cultural legacy as “stimuli to acts of interpretation” (Drucker, 2005), DL initiatives in the humanities must negotiate the need to present the multifaceted nature of cultural heritage while at the same time adhering to the standardizing and streamlining tendencies necessitated by the electronic environment.

This is particularly true for DLs dealing with bi-or multilingual cultural material. Because variation, difference, multilingualism and diachronic change define the environments in which these artifacts emerge, it is necessary to introduce ethnographic, cross-cultural and cross-linguistic perspectives into
the digital representation of cultural objects (Benozzo, 2010). The creators of today’s multilingual humanities DLs must be particularly attune to the challenge facing all global new media: how to avoid stifling linguistic and cultural variety while still providing the most robust, usable, accessible and universal services (Siapera, 2010).

In this paper, we address this larger issue by closely investigating the Belgrade Center for Digital Humanities’ (BCDH) development and implementation of a lexicographic strategy for Serbian cultural heritage DLs. The BCDH is developing its DL projects – “Digital Library of Serbian Cultural Heritage of the 18th Century” and the “Annotated Digital Edition of Vuk Stefanović Karadžić’s Srpski rječnik” – in conjunction with the “Transpoetika Dictionary,” a modular and extensible dictionary service that is seamlessly integrated with the contents of the DLs. As we discuss below, this lexicographic infrastructure establishes the necessary intellectual and technological parameters that allow for implementation of key DL services and features (such as textual mark-up, search and discovery, semantic access etc.), while at the same time creating a rich and flexible environment for the experimentation and play that is at the core of the humanities enterprise.

**Lexicographic Infrastructure: The Dictionary As a Fundamental DL Service**

Today’s DLs are complex online environments whose modular architecture allows integration of various service components. As Greenstein and Thorn (2002) note, the modular structure is “fundamentally liberating” because it makes it possible for libraries to creatively build on services supplied by others and more easily adapt new and evolving technological standards.

Establishing coherent infrastructures has been singled out as an essential requirement for advancing humanities scholarship in the digital age (Borgman, 2009). As installations and services that function as “mediating interfaces,” infrastructures are “structures ‘in between’ that allow things, people and signs to travel across space by means of more or less standardized paths and protocols for conversion or translation.” (Badenoch & Fickers, 2010, p. 11). A digital research infrastructure is no different: it is a mediating set of technologies for research and resource discovery, collaboration, sharing and dissemination of scientific output (see Moulin, Nyhan, Ciula, Kelleher, & Mittler, 2011).

Dictionaries are fundamental tools for humanities scholarship, and as such should form a central part of any research infrastructure. Many important print dictionaries have indeed been digitalized and made available online; web-portals increasingly offer cumulative searches across different dictionaries; and eLexicography is a thriving field (de Schryver, 2003; Fuertes Olivera & Bergenholtz, 2011; Hass, 2005; Hass, 2005; Hockey, 2000; Lemberg, Schröder, & Storrer, 2001; Bergenholtz, Nielsen, & Tarp, 2009). At the same time, however, dictionaries found in todays DLs are still by and large treated as discrete objects: no more and no less than digitalized versions of static print editions. DLs may offer access to electronic dictionaries on their websites, or links to external sites where dictionaries may be consulted, but the kinetics of interacting with a dictionary remains the same as it did in the pre-digital age. A user must interrupt the text, open a new URL, type in the word he or she would like to look up, etc. Today, we go to websites to consult dictionaries – dictionaries do not come to us.

At the BCDH, we contend that rather than existing as a stand-alone separate resource, an electronic dictionary should be deeply embedded in the digital flow of our textual production and reception. This calls for a new consideration of the lexicographic infrastructure as a fundamental component of the DL, as a user-centered lexical reference system that can be queried directly from the text.

Our conception of the role of lexicographic reference materials in DLs is based on the idea of the dictionary not simply as an object but rather as a service; not as a static entity but an interactive method. With a dictionary that is fully integrated into the DL infrastructure via an application programming interface (API), any electronic text becomes a direct entry point to the dictionary. When every word in a DL is a link to a particular dictionary entry, electronic textuality becomes an extension of lexicography: the text begins to contain the dictionary in the same way that the dictionary contains the text.

Previous work in the development of lexicographical infrastructure has included the
proposal for the creation of a federation of heterogeneous dictionary databases (FHDDDB), whose structure would be mapped to a generalized dictionary description language (GDDDL) and queried using a generalized dictionary query language (GDQL) to facilitate the exchange of information between various remotely stored dictionaries in different formats and the end user (Patrick, Zhang, & Artola-Zubillaga, 2000). A human-centered application of Semantic Web technologies lead one author to propose the language grid — an intercultural collaboration experiment that treats existing language services as atomic components and allows users to combine existing components into new services that would be available on the Internet (Calzolari, 2006). Hayashi and Ishida (2006) focus on dictionary access within the proposed language grid, and describe in greater detail a unified abstract dictionary model that can represent a range of machine-readable human dictionaries. Even though the paradigm of distributed and interoperable lexicons has been widely discussed, “little has been made for its practical realisation” (Calzolari, 2006, p. 38). Two more recent initiatives have taken important steps in the direction of standardizing and developing global strategic and methodological visions for building mutually compatible language resources: CLARIN (Common Language Resource and Technology Infrastructure) (Váradi, Wittenburg, Krauwer, Wynne, & Koskenniemi, 2008) and the thematic FLaReNet (Fostering Language Resources Network) (Calzolari, Monachini, & Quochi, 2011). Another fruitful avenue of research has been the application of the Linked Data paradigm (Berners-Lee, 2006) to the publication and representation of linguistic resources on the Web (see Chiarcos, Nordhoff, & Hellmann, 2012).

**Language Culture as the Core of BCDH Projects**

Created against the backdrop of mass-digitization initiatives of the National Library of Serbia, the work of the BCDH represents the trend of smaller, more specialized DLs built by organizations whose institutional independence allows for a greater degree of experimentation and innovation. The BCDH has been developing its DLs with the explicit goal of producing robust, interoperable, user-centered and open-access digital resources for the study and discovery of Serbian linguistic and literary culture. The projects focus on the literary heritage of 18th and 19th century Serbia, and contain key texts such as Dositej Obradović’s Život i priključenija (Life and Adventures) (1793) and Vuk Karadžić’s Srpski rječnik (Serbian Dictionary) (1818, 1852). The digital environment, which allows for both abstract modeling of knowledge as well as its practical representation and multi-level annotation, is ideally suited for exploring the deep valencies of humanistic, particularly ethno-linguistic, trends. Indeed, while these DLs contain digital surrogates and facsimiles of rare and important historical documents, their primary orientation is not geared toward these texts as objects/artifacts. Instead, the BCDH considers the main content of its DLs to be Serbian language itself.

This core concern with language and language culture has defined the intellectual and technological trajectory of the BCDH projects. The language material of 18th-19th century Serbian poses a formidable challenge both to the modern reader and to today’s technologist. While many European languages underwent relatively uninterrupted linguistic evolution, the arena of 18th-19th century Serbian was fraught with dramatic change. The texts included in the “Digital Library of Serbian Cultural Heritage of the 18th Century” reflect the presence of various linguistic layers typical of literary Serbian of the time: 1) the ecclesiastical Serbian recension of Church Slavonic (Serbo-Slavonic); 2) the Russified recension of Church Slavonic known as Ruso-Slavonic; and 3) a fashionable hybrid of Ruso-Slavonic and vernacular Serbian known as Slaveno-Serbian (Albin, 1970). Besides archaic vocabulary and morphology, these 18th century Serbian texts are marked by non-standard textual features (word segmentation, abbreviations, variants of letters and punctuation marks) and varying orthographic conventions.

The progress made by the language reformer Vuk Stefanović Karadžić (1787-1864) to modernize and standardize Serbian in the 19th century is reflected in the “Annotated Digital Edition of Vuk Stefanović Karadžić’s Srpski rječnik.” Though Karadžić eschewed the literary standard(s) of the previous epoch and looked to the vernacular, his dictionary entries contain a multitude of registers, styles, variants and dialects from throughout the Balkan region. Much more than simply an alphabetic list of words with their definitions, the each entry in Karadžić’s Srpski rječnik also provides its counterpart
in German and a Latin, and often supplies ethnographic and anthropological details or examples of Balkan folk storytelling (Ivić, 1990).

The rich and complex linguistic material of 18th–19th century Serbian has encouraged the BCDH to pursue the creation of a digital infrastructure suitable for exploring this content. The digital editions of literary texts in the DLs are deeply marked-up, and thus can offer a broad range of functionalities – such as full-content searching, multiple browsing options, glosses, concordances and indexes, hypertext and hypermedia links. This allows users to define, explain, and order the language and literary culture of various periods of Serbian literary history. These DLs also highlight the inherent contingency and dialogic nature of language systems by inviting users to comment, question, explore, and play with the material through features such as innovative interface and social media.

Transpoetika Dictionary: Toward a Lexicographic Infrastructure

Mindful of general developments in the field of language resources and targeting, and of specific trends in the integration of lexicographic services into DLs, the BCDH has been building a lexicographic infrastructure for the Serbian language from the ground up. This has been a considerable challenge considering the fact that mainstream Serbian lexicography seriously lags behind developments in the field of eLexicography. None of the major national-level initiatives – neither the half-finished, massive Dictionary of the Serbian Literary and Folkloric Language (CAHY, 1959), nor the six-volume Dictionary of the Serbo-Croatian Literary Language (MC, 1967), nor its bulky, oversized one-volume counterpart (MC, 2007) – have been made available in electronic format, either on- or offline. The computational resources being developed at the University of Belgrade (Krstev, Vitas, & Pavlović-Lažetić, 2008; Obradović & Stanković, 2008) have not been used to produce human-oriented dictionaries.

The core infrastructural component of the BCDH projects and the basis for the Serbian lexicographic infrastructure is the “Transpoetika Dictionary,” a collaborative, web-based, multi-media, scalable and interoperable, bilingualized Serbian-English dictionary aligned with Princeton University’s WordNet and geared towards both readers and learners of Serbian. (Tasovac, 2009; Tasovac, in press).

Like WordNet, Transpoetika overcomes the limitations of the print format by ordering lexemes based on conceptual affinity (rather than abstract, alphabetic order) and relational complexity through hyponyms, hyponyms etc. Transpoetika is based on a corpus of nearly 50 million words, which is expanded almost daily with new texts obtained both online and offline. The dictionary can be accessed via a Web interface and as a web service.

Web Interface and Interaction With Other Web Services

Through its Web interface, Transpoetika can actively participate in the creative re-mixing of data within the network architecture of data sharing, usually referred to as Web 2.0 (O’Reilly 2007). In addition to providing classical dictionary information (lemma, part of speech, monolingual definitions, translations into English) and information on WordNet-specific lexical and semantic relations between entries, each word in Transpoetika, for instance, contains a so-called ‘Live Quote’ link: when the dictionary user clicks on it, the system retrieves the latest examples of the entry word from Twitter (http://twitter.com) — the popular social network and microblogging platform where users post status updates and communicate using short messages limited to 140 characters. These quotes are displayed directly within the dictionary entry. Pronunciation of entries is provided by Forvo (http://www.forvo.com), a web-based platform for recording and listening to audio files of individual words and short phrases in more than 200 languages.

Transpoetika uses the external photo-sharing platform Flickr (http://www.flickr.com) for displaying images related to dictionary entries. While Flickr users upload and store photographs on the web, they can also annotate them, using freely chosen tags or keywords. This so-called social tagging system does not rely on a controlled vocabulary, but rather on “shared and emergent social structures and behaviors, as well as related conceptual and linguistic structures of the user community” (Marlow et al. 2006, p. 31). Each tag serves as a link to other images tagged with the same keyword, and the Flickr API makes it possible to retrieve images based on their tags and to display them on a different website.
Transpoetika as a Web Service

The contents of the Transpoetika Dictionary are accessible via a RESTful web API. The service returns results in both XML and JSON formats and lets users query the dictionary for definitions and/or English translations of a specific word.

A separate API call searches our Serbian MorphoSyntax database (SMS), which contains all the forms of all the dictionary lemmas marked with an appropriate tag indicating its classificatory and inflectional features. The tagset we are using is a modified version of the MULTEXT-East tagset for Serbian (Erjavec, 2009).

The Transpoetika API is at the moment not publicly available due to budgetary constraints. But it is actively used within various projects at BCDH. API calls to the SMS database make it possible to search for Transpoetika Dictionary entries in their oblique forms, a very important feature for a morphologically rich language like Serbian. The user of the dictionary web interface can also retrieve all the forms related to an entry and display full declensions or conjugations with a click of a mouse.

Digital Library of Serbian Cultural Heritage of the 18th Century

The goal of the “Digital Library of Serbian Cultural Heritage of the 18th Century” is to allow today’s users to meaningfully explore and interact with texts of the period known as the Serbian and South Slavic Age of Enlightenment. To help users navigate these lexically and orthographically complex documents, we have created a hybrid edition where texts are doubly-encoded: once using traditional 18th-century orthography (including the graphemes that have since disappeared from Serbian) and again using modernized and standardized Serbian spelling rules. The corpus contains linguistic and semantic annotations that add modern phonetic, morphological, lexical and conceptual equivalents to the largely archaic vocabulary. These are linked to and provided by the Transpoetika Dictionary API. Users are able to search the contents using both modern and 18th-century spellings.
The Annotated Digital Edition of Vuk Stefanović Karadžić’s Srpski rječnik

Vuk Stefanović Karadžić, the linguist, folklorist and reformer of the Serbian language, published his landmark Srpski rječnik, istumačen njemačkijem i latinskijem riječima — Lexicon Serbico-Germanico-Latinum in (1818 and 1852) as the first lexicon of the modern Serbian vernacular. The dictionary has a unique place not only in the history of Serbian, but the South Slavic diasystem in general. However, since entries are written mainly in a dialect on the margins of contemporary standard Serbian, they are not always easily understood by contemporary speakers. To make the Srpski rječnik more usable for today’s scholar and student, we are marking-up and annotating structural elements (lemma, part of speech, senses, definitions, translation equivalents, examples etc.), and supplying identification tags for other key features, such as: Serbian equivalents for dialect word forms, Serbian ekavian word-forms to both standard and east-Herzegovina jekavian entries, labeling of Turkisms overlooked by previous editors, etc. Furthermore, we assign semantic domain labels to word senses in accordance with Magnini and Cavaglia (2000) and Bentivogli et al. (2004). All entries and semantic domain labels are cross-referenced with the Transpoetika Dictionary.

A significant feature of this digital edition is the integration of social media features to enhance user interaction with the text. We publish individual Srpski rječnik entries in regular intervals on our dedicated website “Reklakaza.la” (Serbian for “hearsay”), and link them via Twitter (http://twitter.com/Vuk_Karadzic) and Facebook (http://facebook.com/reklakaza.la). As attested by our nearly 25,000 “friends” on Facebook and more than 300 Twitter followers, we have discovered a broad audience outside highly professionalized linguistic circles that is both curious and enthusiastic about exploring the historic and ethnographic fabric of a language (Kirkness, 2008). Indeed, these social media platforms have proven to be ideal places for sharing and showcasing the Srpski rječnik, and for generating meaningful dialogue around it. As short, yet richly complex units of linguistic and cultural material, Srpski rječnik entries meet both Facebook and Twitter’s space limits and can be posted and circulated through re-posting and re-tweeting. All posted entries generate lively commentary by users throughout the world, who bring both intellectual and personal reflection, critique and inquiry to the texts. Encountering Karadžić’s dictionary in the context of today’s most popular websites integrates cultural heritage into users’ daily informational newsfeeds, and helps decrease the distance between the historical text and the modern reader. The DL becomes platform for bringing meaningful humanities inquiry into the public conversation, and fosters a sense of community, sharing, and mutual learning.
Conclusion
As powerful tools for learning and research, “Transpoetika,” the “Digital Library of Serbian Cultural Heritage of the 18th Century” and the “Annotated Digital Edition of Vuk Stefanović Karadžić’s Srpski rječnik” offer new ways for scholars and students to explore various aspects of the Serbian linguistic and literary fabric. As digital representations of cultural heritage and as infrastructural web-services, these projects do not simply restate established narratives of Serbian cultural history, as is often the case with large institutionally-created DL projects (Dalbello, 2009). Rather, they open up new lines of communication between the modern reader and historical texts, creating new fields of interpretive participation where cultural heritage is not simply displayed as defined, objectified artifacts, but rather as complex living texts that can be continually contextualized and re-contextualized. The APIs for the Transpoetika Dictionary and the Serbian MorphoSyntax database form the foundation of an emerging lexicographic infrastructure for the Serbian language that can be used to enhance digital editions of Serbian texts and create to user-friendly, semantically rich DLs.

REFERENCES
In Granger, Sylviane and Magali Paquot (eds.), Electronic Lexicography, Oxford: Oxford University Press.