APIs in Digital Humanities: The Infrastructural Turn

Toma Tasovac, Adrien Barbaresi, Thibault Clérice, Jennifer Edmond, Natalia Ermolaev, Vicky Garnett, Clifford Wulfman

APIs in Digital Humanities: The Infrastructural Turn

Toma Tasovac
ttasovac@humanistika.org
Belgrade Center for Digtial Humanities

Adrien Barbaresi
adrien.barbaresi@oeaw.ac.at
Austrian Academy of Sciences

Thibault Clérice
thibault.clerice@uni-leipzig.de
University of Leipzig

Jennifer Edmond
jedmond36@gmail.com
Trinity College Dublin

Natalia Ermolaev
nataliae@Princeton.EDU
Princeton University

Vicky Garnett
garnetv@tcd.ie
Trinity College Dublin

Clifford Wulfman
cwulfman@Princeton.EDU
Princeton University

As a community of practice, digital humanists deal with data and metadata not as static artifacts, but rather as complex, multi-dimensional and multi-layered datasets that can be analyzed, annotated and manipulated in order

to produce new knowledge. One of the most important challenges facing DH today is how to consolidate and repurpose available tools; how to create reusable but flexible workflows; and, ultimately, how to integrate and disseminate knowledge, instead of merely capturing it and encapsulating it. This technical and intellectual shift can be seen as the "infrastructural turn" in digital humanities (Tasovac et al. 2015).

Application Programming Interfaces (APIs) have the potential to be powerful, practical building blocks of digital humanities infrastructures. On the technical level, they let heterogeneous agents dynamically access and reuse the same sets of data and standardized workflows. On the social level, they help overcome the problem of "shy data", i.e. data you can "meet in public places but you can't take home with you" (Cooper 2010). Some 10 years ago, Dan Cohen started the conversation about APIs in DH by pointing out that, despite their potential, Andreas few humanities projects — in contrast to those in the sciences and commercial realms — were developing APIs for their resources and tools (Cohen 2005). In the decade since, API development in the digital humanities has certainly increased: today, both large-scale, national and international initiatives, such as HathiTrust, DPLA or Europeana, as well as individual projects, such as Canonical Text Services (CTS), Open Siddur, Folger Digital Texts, correspSearch etc., are focusing their attention and resources on developing APIs. It is now time to reflect on this development: have standards or best-practices evolved? What workflows are most effective and efficient for creating APIs? What are the challenges or stumbling blocks for creating or using APIs? Are APIs being used by DH researchers? What is the future of API development and use in the humanities community?

This panel will cover both the theory and practice of APIs in the digital humanities today. It will bring together researchers working on major European and North American projects, who will discuss APIs from the perspectives of design, implementation, and use, as well as technical and social challenges. Each group will have 10 minutes for their statement, and 40 minutes will remain for group discussion and questions from the audience. One of the panel members will serve as the moderator. All speakers have confirmed their intention to participate in the panel.

Toma Tasovac (Belgrade Centre for Digital Humanities) will discuss an API-centric approach to designing and implementing digital editions. Starting with the notion of text-as-service and textual resources as dynamic components in a virtual knowledge space, Tasovac will show how two recent projects — *Raskovnik: A Serbian Dictionary Platform* and *Izdanak: A Platform for Digital Editions of Serbian Texts* — were implemented using API-focused data modeling at the core of the project design process. The API-first approach to creating TEI-encoded digital editions offers tangible interfaces to textual data that can be used in tailor-made workflows by humanities researchers and other users, well-suited to distant reading techniques, statistical analysis and computer-assisted semantic annotation. The "infrastructural turn" in Digital Humanities does not only have practical implications for the way we build tools and create resources, but also has theoretical ramifications for the way we distinguish highly from loosely structured data: if text is not an object, but a service; and not a static entity, but an interactive method with clearly and uniquely addressable components, a formal distinction between a dictionary and, say, a novel or a poem, is more difficult to maintain.

Clifford Wulfman and **Natalia Ermolaev** (Center for Digital Humanities, Princeton) will discuss the design and implementation of Blue Mountain Springs, the API for the Blue Mountain Project's collection of historic avant-garde periodicals. By modeling magazine data using the FRBRoo ontology and its periodical-oriented extension PRESSoo (PRESSoo, 2014), this RESTful API exposes the Blue Mountain resource in a variety of data formats (structured metadata, full-text, image, linked data). The authors will provide several examples of how Blue Mountain Springs has been used by researchers, drawing especially from the results of the hackathon they will host at Princeton in February 2016, which will bring together approximately twenty periodical studies scholars, technologists, and librarians to work with the API. Creating APIs is part of a trend in DH to move into a post-digital-library phase, when the traditional library functions of discovery and access are no longer sufficient to support research in the humanities. This trend also suggests that DH researchers must reconceptualize their own engagement with material, to think less in terms of monographs and more in terms of resources, and consequently to promulgate their work not as web sites but as web services.

Thibault Clérice (University of Leipzig) will discuss the design of the Canonical Text Services (CTS) and its URN scheme, which make the traditional citation system used by classicists machine-actionable (Blackwell and Smith 2014)[1]. The Homer Multitext (HMT) implementation of CTS requires textual data to be extracted out of its original digital representation into RDF triples in order to be served. The Perseus Digital Library (PDL) implementation, on the other hand, uses extended transformations to slice XML files into multiple records, each representing a passage at a certain level. While relational and RDF database approaches have had some success in scalability and speed (Tiepmar 2015), they also have to deal with maintenance and evolution capacity. There is a real need for this type of DH projects to scale not only in terms of data retrieval speeds, but also in terms of allowing researchers to correct and enhance their data. In addition, projects need to be able to propose other narratives: sliced data doesn't easily provide access to the full data model. Clérice will discuss

why and how, using both a native XML-based system such as eXist and a Python-based implementation, one can achieve scalability while guaranteeing maintenance and evolution.

Adrien Barbaresi from the Austrian Academy of Sciences (ICLTT) will discuss the use of APIs in building resources for linguistic studies. The first case deals with lesser-known social networks (Barbaresi 2013) while the second tackles the role of the Twitter API in building the ICLTT's "tweets made in Austria" corpus[2]. For computational linguists, short messages published on social networks constitute a "frontier" area due to their dissimilarity with existing corpora (Lui & Baldwin 2014), most notably with reference corpora of written language. Since data are mainly accessed and collected through APIs and not in the form of web pages, Barbaresi argues that social networks are a frontier area for (web) corpus construction. He will point out the challenges of using Twitter's API, for example how to reveal the implicit decisions and methodology used by API designers, as well as concrete implementation issues, such as the assessment and optimization of data returned by the API. Free APIs may come at no cost, but they also offer no guarantee, so that the use of commercial APIs for research purposes has to be seen with a critical eye in order to turn a data collection process into a proper corpus.

Finally, **Jennifer Edmond** and **Vicky Garnett** (Trinity College Dublin), will provide reflections on the place of APIs within European research infrastructures for the humanities. Their contribution to the panel builds on their recent study on the Europeana Cloud project, which found that while access to data is a real and growing area of interest, very few humanities researchers seem to actively and directly use APIs.[3] They will describe two initiatives, one technical, one social, aiming to better harness the potential of the API to meet researcher's implicit needs. The first is the Collaborative European Digital Archival Research Infrastructure (CENDARI) project, whose platform is structured around an internal API that will allow multiple data sources (local repository, triple store, metasearch engine) to be aligned, enhanced and then served out to a number of environments and tools, including the project's native note-taking environment. The second example is the genesis and development of the concept of the 'inside-out' archive. This framework, which has arisen out of a collaborative venture between several European humanities research infrastructure projects, seeks to encourage collection holding institutions to look beyond their own digitization programs and platforms and recognize the rising importance of machines-as-users (requiring specific access points and formats) rather than the somewhat outdated model of individual institutional web presence serving individual human resource seekers.

The five speakers on this panel will address some of the most pressing issues related to the ongoing development and future of APIs on the DH research infrastructure landscape. The discussion will cover both micro- and macro levels, ranging from methodological implications and technical scalability to the ways in which API-based data access to collections challenges traditional norms of institutional identity and independence. As such, the panel will offer a timely platform for a multifaceted debate on the potentials and pitfalls of building and using APIs in the digital humanities.

Bibliography

Badenoch, A. and Fickers A. (2010). Europe Materializing? Toward a Transnational History of European Infrastructures. In Badenoch, A. and Fickers A. (eds.), Materializing Europe: Transnational Infrastructures and the Project of Europe, 1-26. Basingstoke, Hampshire; New York: Palgrave Macmillan.

Barbaresi, A. (2013). Crawling microblogging services to gather language-classified URLs. Workflow and case study. In Proceedings of the 51th Annual Meeting of the ACL, Student Research Workshop, pages 9-15.

Blackwell, C. and Smith N. (2014). *Canonical Text Services Protocol Specification*. http://folio.furman.edu/projects/citedocs/cts/. Accessed October 23, 2015.

Cohen, D. Do APIs Have a Place in the Digital Humanities. http://www.dancohen.org/blog/posts/do_apis_have_a_place_in_the_digital_humanities. Accessed October 24, 2015.

Cohen, D. (2006). "From Babel to Knowledge: Data Mining Large Digital Collections." *D-Lib Magazine* 12, no. 3.

Cooper, D. (2010). When Nice People Won't Share: Shy Data, Web APIs, and Beyond, *Second International Conference on Global Interoperability for Language Resources (ICGL 2010)*, n. pag.

Edmond, J, Bulatovic N. and O'Connor, A. "The Taste of 'Data Soup' and the Creation of a Pipeline for Transnational Historical Research." Journal of the Japanese Association for Digital Humanities 1, no. 1 (2015): 107–22.

Edmond, J. and Garnett V. (2015). APIs and Researchers: The Emperor's New Clothes, International Journal of Digital Curation, 10(1): 287-97.

LeBeuf, Patrick (ed.). PRESSoo. Extension of CIDOC CRM and FRBROO for the modelling of bibliographic information pertaining to continuing resources. Version 0.5, http://www.ifla.org/files/assets/cataloguing/frbr/pressoo_v0.5.pdf. Accessed, November 1, 2015.

Murdock, J. and Allen C. (2011). InPhO for All: Why APIs Matter, Journal of the Chicago Colloquium on Digital Humanities and Computer Science 1(3): http://www.jamram.net/docs/jdhcs11-paper.pdf. Accessed, October 23, 2015.

Tasovac, T. Rudan S. and Rudan S. (2015). Developing MorphoSLaWS: An API for the Morphosyntactic Annotation of the Serbian Language. Systems and Frameworks for Computational Morphology, 137-47. Heidelberg: Springer.

Tiepmar, J. (2015). Release of the MySQL based implementation of the CTS protocol. In Bański, P., Biber H., Breiteneder E., Kupietz M., Lüngen H. and Witt A. (eds.), Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3), 35-43. Mannheim: Institut für Deutsche Sprache.

Notes

[1] Infrastructures are installations and services that function as "mediating interfaces" or "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 and Fickers 2010, 11). Digital research infrastructures are no different: they are a mediating set of technologies for research and resource discovery, collaboration, sharing and dissemination of scholarly output.

[2] http://bluemountain.princeton.edu

[3] Traditional scholars have been citing texts the same way for centuries : for example, "Hom. Il. 1.1", which corresponds to Homer's Iliad, Book 1 Line 1. However, although the passage identifier does not change from one scholar to another in most cases, the abbreviation used for the author and the work title will diverge among authors, countries, and publications.

[4] http://www.oeaw.ac.at/icltt/node/193

[5] This is a result of the confluence of several factors: that the research data most humanists want to access is seldom available via an API; that many sources that do offer relevant data lack the structure or metadata to make the API useful; and that humanistic researcher generally lack the skill set to experiment directly with APIs. At the same time, however, one of the key desires expressed by researchers in the current landscape is the federation of high-quality data (Edmond and Garnett 2015).