Stanford | LIBRARIES

Socially impactful digital libraries and/as challenges to NLP application

Mark A. Matienzo / **@anarchivist**
Stanford University Libraries
nlp4arc, Chapel Hill, NC, 2 February 2018 / **#nlp4arc**

Understanding social impact

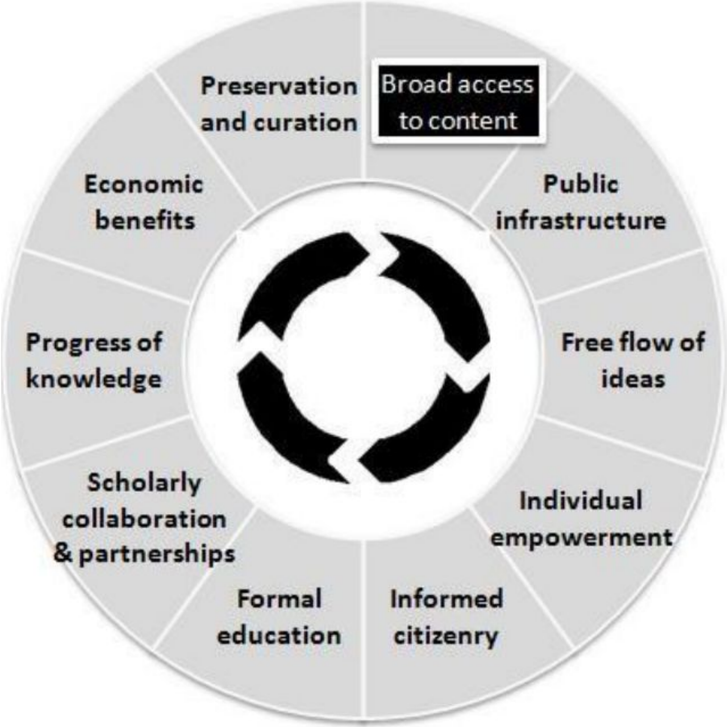
- “the net effect of an activity on a community and the well-being of individuals and families” (Centre for Social Impact)
- “a significant, positive change that addresses a pressing social challenge” (University of Michigan Ross Center for Social Impact)

Ethics, social impact, and NLP

- Small but growing area of research
- Hovy and Spruit 2016 provide pivotal framing
  - Impact of NLP on social justice
  - Language as proxy for human behavior
  - Negative impact factors: *exclusion, overgeneralization, bias confirmation, topic overexposure,* and *dual use*

Social impact and digital libraries

- Long history in LIS literature re: libraries and social impact (Poll and Payne 2006, Oakleaf 2010, Kerslake and Kinnell 1998)
- DLs as platforms for humanitarian information (Witten 2005)
- Frameworks for social impact of DLs: Tanner and Deegan 2011, Calhoun 2014
- ***Socially impactful digital libraries* have a focus to improve society**
- Emphasizes libraries as non-neutral actors (cf. Bourg 2015)



Calhoun’s framework of DL social roles

Socially impactful digital libraries at Stanford

- Portfolio development as a strategic goal; occasionally referred to as “humanitarian data”
- Broader context of similar projects with loose connection both within Stanford Libraries and across the university
- Two specific projects for this presentation
    - **Digital Library of the Middle East**
    - **Virtual Tribunals**

Digital Library of the Middle East

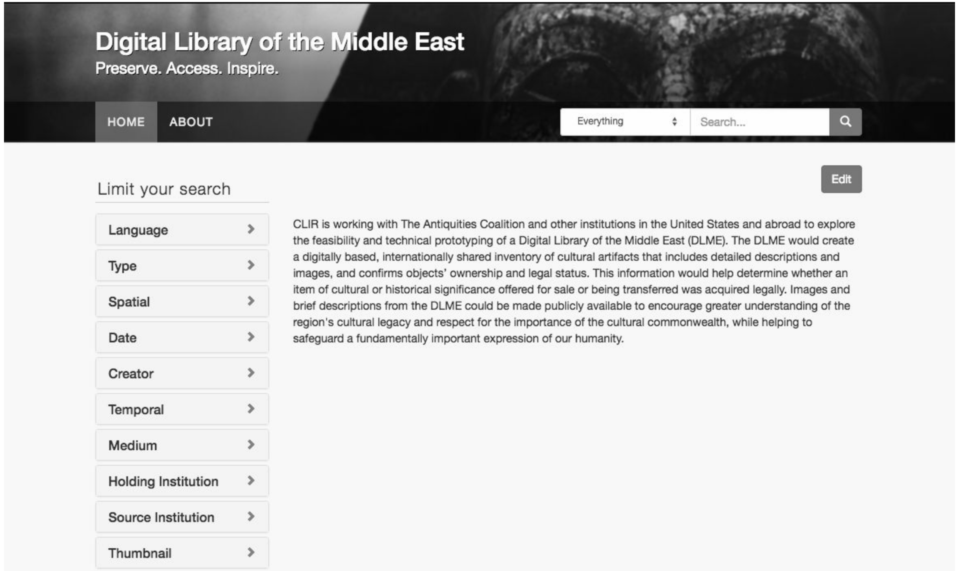
https://dlme.clir.org/

- Partnership between the Council on Library and Information Resources and the Antiquities Coalition, with Stanford Libraries as technical partner
- Response to conflict and loss of life in Middle East and increased looting of and trafficking in cultural heritage materials
- Broad and expansive vision outlined to stakeholders and funders
- Stanford component focuses on providing discovery platform for cultural heritage material from the Middle East and North Africa

DLME prototype and data challenges

https://spotlight.dlme.clir.org/

- Aggregated metadata from 7 institutions (135K+ items)
- Common entities but sparse records, differing practices
- Variants in vocabulary, transliteration, language



CLIR is working with The Antiquities Coalition and other institutions in the United States and abroad to explore the feasibility and technical prototyping of a Digital Library of the Middle East (DLME). The DLME would create a digitally based, internationally shared inventory of cultural artifacts that includes detailed descriptions and images, and confirms objects' ownership and legal status. This information would help determine whether an item of cultural or historical significance offered for sale or being transferred was acquired legally. Images and brief descriptions from the DLME could be made publicly available to encourage greater understanding of the region's cultural legacy and respect for the importance of the cultural commonwealth, while helping to safeguard a fundamentally important expression of our humanity.

Stanford | LIBRARIES

Virtual Tribunals

http://stanford.io/2DVwtAh

- Collaboration between Stanford Libraries and the WSD Handa Center for Human Rights and International Justice
- Goal is to build a comprehensive digital library platform for materials from international criminal tribunals and truth and reconciliation commissions
- Current project focusing on materials from East Timor tribunal in four languages (English, Portuguese, Indonesian, Tetun)
- Planned expansion into other tribunals (means other languages)
- Entities include people, places, types of crimes/conflict, and many cross-references and citations

Challenges: underresourced languages

- Underresourced languages in an NLP sense are those which are missing machine-readable language data
- Not all underresourced languages are “endangered”
    - Consider Arabic (El-Haj 2015) and Sorani Kurdish (Walther and Sagot 2010)
- “All ASEAN languages are underresourced” (Ye 2016)
- Hovy and Spruit 2016 acknowledge commercial incentive to focus on overexposed languages (e.g. English)

Challenges: consistency

- Variation in transliteration schemes, orthography, etc. across communities represented in data
- Use of LC transliteration in subject headings, place names, etc. may not reflect more familiar or common transliteration

Challenges: entity extraction, reconciliation, etc.

- Particularly a challenge for underresourced languages
- Specialized concepts (e.g. projects like Virtual Tribunals)
- Impact of concept reconciliation as vector for introduction of bias or oppressive language

Challenges: avoiding reproduction of bias

- Hovy and Spruit 2016 identify introduction of bias as a major factor for production of linguistic data sets
- With regard to specific projects mentioned, major concern is how to avoid or eliminate Eurocentric and imperialistic biases
- Thought experiment: **What is the social impact of using the Bible to produce corpora for a less-resourced language?** ...especially if that community was impacted by Christian imperialism?

Challenges: ethical tool selection

- Promising tools exist that have been developed or funded in service of military and intelligence needs
- Anderson et al. 2012 identifies DARPA workshops as critical to development of computational linguistics research
- Hovy and Spruit 2016 consider government and military involvement in computational linguistics be worth a closer study
- **How do we help LIS/archives practitioners to make informed choices?**

References

- Anderson, Ashton, Dan McFarland, and Dan Jurafsky. 2012. “Towards a Computational History of the ACL: 1980-2008.” In *Proc. ACL Workshop on Rediscovering 50 Years of Discoveries, 13–21*.
- Bourg, Chris. 2015. “Never Neutral: Libraries, Technology, and Inclusion.” OLA Superconference, Toronto.
- Calhoun, Karen. 2014. “Social Roles of Digital Libraries.” In *Exploring Digital Libraries: Foundations, Practice, Prospects.* Neal Schuman.
- El-Haj, Mahmoud, Udo Kruschwitz, and Chris Fox. 2015. “Creating Language Resources for Under-Resourced Languages: Methodologies, and Experiments with Arabic.” *Language Resources and Evaluation* 49 (3): 549–580.
- Hovy, Dirk, and Shannon L. Spruit. 2016. “The Social Impact of Natural Language Processing.” *Proc. 54th ACL*, 2:591–598.
- Kerslake, Evelyn, and Margaret Kinnell. 1998. “Public Libraries, Public Interest and the Information Society: Theoretical Issues in the Social Impact of Public Libraries.” *J. Librarianship and Information Science* 30 (3): 159–67.
- Oakleaf, Megan. 2010. *The Value of Academic Libraries: A Comprehensive Research Review and Report.* ACRL.
- Poll, Roswitha, and Philip Payne. 2006. “Impact Measures for Libraries and Information Services.” *Library Hi Tech* 24 (4): 547–562.
- Tanner, Simon, and Marilyn Deegan. 2010. *Inspiring Research, Inspiring Scholarship.* JISC.
- Walther, Géraldine, and Benoît Sagot. 2010. “Developing a Large-Scale Lexicon for a Less-Resourced Language: General Methodology and Preliminary Experiments on Sorani Kurdish.” In *Proc. 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages.*
- Witten, Ian H. 2005. “Digital Libraries and Society: New Perspectives on Information and Dissemination.” In *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific*, 191–215. IGI Global.
- Ye Kyaw Thu. 2016. “Challenges of Natural Language Processing Research for Under-Resourced Languages, using Myanmar Language as an Example.” Conf. on Khmer NLP, Phnom Penh.

Thank you!

Stanford | LIBRARIES

Mark A. Matienzo / **@anarchivist**
Stanford Libraries
nlp4arc, 2 February 2018