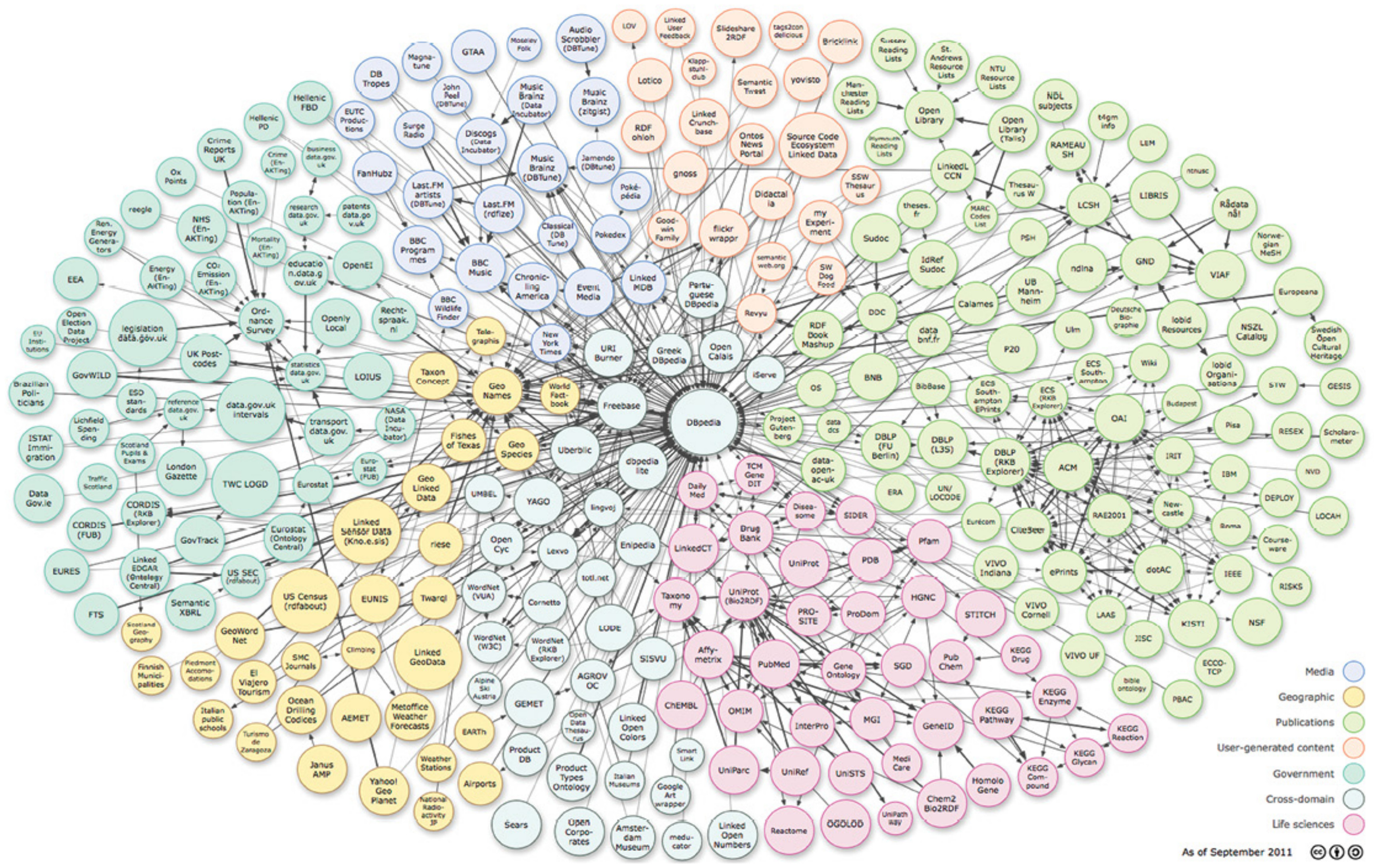
Yes, but is it *Linked* Open Data?

The Question:

Developed in libraries, linked open data has found dedicated advocates in the digital humanities, aided by community values that commit to openness and collaboration as well as the relatively constrained nature of many projects. But has this advocacy extended to application? And do those digital humanities projects that promise linked open data deliver? This poster presents preliminary results from an analysis of 407 successful proposals to the National Endowment for the Humanities' Office of Digital Humanities grant program. Our classification of funded projects according to their commitment to produce or apply linked and/or open data reveals a strikingly small subset. In our next stage of analysis, we will determine whether those few projects that purport to produce linked open data actually develop and contribute this data. Our goal is to provide data on actual rather than anecdotal uptake of linked open data in funded digital humanities projects, and to recognize potential barriers and confusion around ambition and implementation of this emergent technology.



The Semantic Web: Linked open data cloud diagram as of September 2011 by Anja Jentzsch.
CC-SA 3.0 via Wikimedia Commons.

Background:

Linked open data is structured information that is published on the web, can be queried contextually by machines, and is openly accessible to the public.

First coined in 2006 by Tim Berners-Lee, linked data builds on existing web technologies to identify, then connect distinct pieces of information contextually. By structuring information according to this model, linked open data becomes the building block for the semantic web (Web 2.0), which allows institutions, projects, and platforms to share and reuse information. Most importantly, Web 2.0 resources accrue authority through use and reference, which enhances the discoverability of robust, accurate information.

The building blocks of linked open data include Unique Resource Identifiers (URIs), which are published data points existing on the web, and standardized vocabularies that can interlink, define, and give context to relationships between URIs.

project
digital
humanities
development
data
scholars
research
university
open
history
web
tools
source
online
based
using
new
workshop
software
library

Cirrus visualization of 407 NEH-ODH Project Descriptions. "Data" occurs 273 times in the corpus.
QR Code links to corpus in Voyant.

Open Data

is "freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control."

Wikipedia

Linked Data

is "a method of publishing structured data so that it can be interlinked and become more useful through semantic queries. It builds upon standard Web technologies … to share information in a way that can be read automatically by computers."

Wikipedia

So Linked Open Data is *both* Linked and Open.

Elizabeth Grumbach
@EMGrumbach
Research Associate
Initiative for Digital Humanities, Media & Culture
Texas A&M University

Spencer D. C. Keralis
@hauntologist
Head, Digital Humanities & Collaborative Programs Unit
University Libraries
University of North Texas

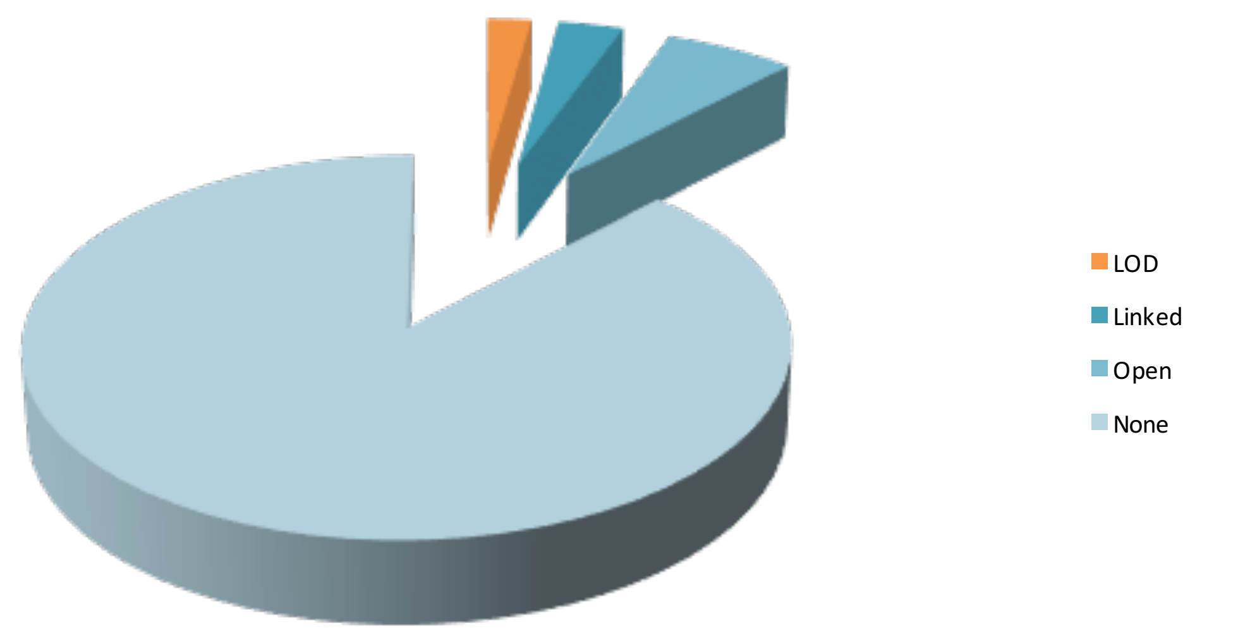
Sarah Potvin
@sp_meta
Digital Scholarship Librarian
University Libraries
Texas A&M University

Methods:

To develop the corpus, we searched the NEH Funded Projects Query Form for "Digital Humanities" in the "Division or Office" field. This generated a list of 407 projects funded by the NEH-ODH. We conducted a preliminary analysis of all Project Descriptions in Voyant to determine overall patterns among the corpus and begin tightening our focus. The Cirrus word cloud of the complete corpus is at left with a QR code for the corpus. Interestingly, "data" occurs 273 times in the corpus. If data is only used once in each project description, this suggests only slightly more than half of the projects have data as a foci in their Project Descriptions (the percentage is lower, since those projects referring to data tend to use it more than once). "Open" appears 182 times, so potentially in fewer than 45% of the projects.

We then isolated 144 Project Descriptions from the corpus by doing keyword searches for "linked," "open," and "semantic," in an attempt to narrow the corpus to those projects most likely to indicate Linked Open Data. We manually coded these Project Descriptions to indicate either **LOD** (linked open data), **Open** (data open or shareable), or **Linked** (linked within a project or between projects, but not necessarily open), or **None** (no indication of linked or open data derived from project).

determine what scholars actually intend to do with their data. A significant number of projects do appear to rely on or aim to produce Open Source software. While laudable, Open Source development work is outside the scope of the present discussion. Further content analysis is necessary to ensure robust intercoder reliability, as well as to apply more granular classifications to identify supported projects that, while foundational to the development of LOD standards, systems, and platforms, are not aimed at actual data production.



Percentages of NEH-OHD Funded Projects indicating Linked Open Data (2%), Open Data (6%), Linked within or between projects (12%), or no indication of linked or open data (80%)

Analysis:

Based on our coding, only 8 projects are producing Linked Open Data; an additional 12 are using or producing data that is linked, either within the project or to data produced by another project, but not necessarily open; 24 other projects are producing data that is at least nominally open but not explicitly LOD. Of the remaining 100 projects, some have no relationship to LOD at all; others, while making no claim to produce linked open data, linked data, or open data, are aimed at the development of platforms or processes for storing, exchanging, and maintaining these types of data. This means that, according to our preliminary coding, of the 407 projects funded by the NEH-ODH since its inception in 2008, only 1.9% are explicitly producing Linked Open Data.

Further, despite the NEH-ODH's 2012 mandate for funded projects to include plans for sharing project data, only 7.8% of funded projects make clear that open data is an explicit product of the project. Open data does not seem to be a priority in most projects.

However, there is a significant amount of conflation between terms in the Project Descriptions. For example "Open Source," which typically describes code or software openly available for reuse and repurposing, is used interchangeably within the corpus with the terms "Open Access" or just "Open" in describing data, databases, or other resources. This fluidity, or indeed, misuse, of terminology makes it difficult to accurately

More Questions:

Unsurprisingly, this preliminary analysis generates more questions than it answers:

- Are funding agency priorities driving the development of semantically isolated projects by privileging the development of stand-alone tools?
- Are other funding agencies or foundations in the U.S. prioritizing the development of semantic web projects and technologies?
- Does the lack of LOD in U.S. DH demonstrably damage these projects by impairing discoverability, and limiting access and authority?
- Why is Linked Open Data a catchword if so few projects in the (U.S.) Digital Humanities rely on LD or LOD technologies or standards?

Acknowledgments:

Attendees at the "Linked Data & Digital Humanities" panel at the Texas Library Association 2016 conference generated interesting questions, and we're grateful to Jeanette Sewell of Houston Public Library for organizing it. Laura Schadt, Collaborative Programs Student Assistant with the UNT Libraries, contributed significantly to compiling the data for this project. Ann Sneesby-Koch at the NEH Office of Digital Humanities was generous with her assistance.