Digital Humanities Workshop – RDA/US Meeting Reporting

Meeting: May 28, 2015

Location: John Hopkins University, Baltimore, MD

Organizing Committee: Sayeed Choudhury, Bridget Almas, Larry Lannom, Kim Fortun, Mike Fortun, Candice Lanius

Report Prepared By: Candice Lanius

This workshop was sponsored by the Research Data Alliance (RDA/US) to learn more about the disciplinary needs of digital humanists in regards to data infrastructure and data sharing. Representatives of RDA/US were present to share the latest information regarding global infrastructure development efforts, scholars from several digital humanities projects provided insight into their needs through use cases, and funders were available to comment on the economics and sustainability of digital humanities data infrastructure.

Attendees:

Name	**Organization**		**Name**	**Organization**
Bridget Almas	Perseus Digital Library		Larry Lannom	CNRI and RDA/US
Sayeed Choudhury	Johns Hopkins		Wendy Queen	Johns Hopkins University Press
Brett Bobley	NEH		Tom Elliot	ISAW
Fran Berman	RPI and RDA/US		Candice Lanius	RPI and RDA/US
Greg Crane	Perseus Digital Library		Beth Plale	IU and RDA/US
Helen Cullyer	Mellon		Stephanie Simms	UCLA
Kim Fortun	RPI - PECE		Trevor Owens	IMLS
Lindsay Poirier	RPI - PECE		Chris Blackwell	Furman University
Luis Felipe Murillo	RPI - PECE		Marie-Claire Beaulieu	Tufts University
Mark Patton	Johns Hopkins		Bruce Robertson	Mount Allison University
Carsten Thiel	Goettingen University/ DARIAH			

Agenda:

10:00 – 10:15	Welcome from Fran Berman and Introductions
10:15 - 10:30	Overview of Day
10:30 - 11:00	Presentation of RDA Infrastructure (Larry Lanom, Kim Fortun)
11:30 – 12:00	Discussion of Use Cases in relation to RDA Infrastructure (Bridget Almas, Sayeed Choudhury)

1:00 - 2:00	Reactions, Comments, Etc. from Workshop Participants
2:30 - 3:30	Further Discussion and Feedback from RDA/US
3:30 - 4:00	Next Steps

Welcome from Fran Berman and Introductions

Fran Berman began the meeting by offering a question to the attendees: What can RDA do for you? This became the mantra of the day, finding ways to utilize RDA's expertise and international community space to support the particular needs of digital humanities projects. To initiate this discussion, Fran explained what RDA has to offer. Data sharing is necessary in order for empirical disciplines to advance and seize opportunities for new research. RDA provides a space that prioritizes developing and implementing the necessary infrastructure for data sharing, whether that infrastructure is technical or social in nature. Infrastructure is developed through community efforts in the form of interest groups and working groups, and any outputs must be made *for* someone to address a practical problem and be generalizable for a community of practice. Fran then explored how RDA can assist with building DH infrastructure. First, RDA already has existing infrastructure that can be tailored to the needs of the DH community, and RDA provides coordination and collaboration to develop new infrastructure. Second, RDA has a large, international community that is well networked to share and work with a large group of experts. And third, RDA works to empower students and early career professionals in their specific fields. All workshop participants were invited to join the RDA by agreeing to the organization's core principles and signing up for a free account on www.rd-alliance.org.

Overview of the Day – Sayeed Choudhury

Sayeed next provided a brief overview of the day's events. During this time, he offered a few comments on what infrastructure means for the digital humanities. First, infrastructure is not a singular noun; “infrastructure” must be understood as “infrastructures.” Additionally, there is an interpretive layer for data that means it is not analogous to other utilities. That is, a water pipeline does not interpret water, but data sharing capabilities do have an effect on the data's meaning. Sayeed also discussed the need to create funding incentives for projects to build or reuse interoperable and sustainable infrastructure. A few workshop participants had complicating comments to add about infrastructure:

- For an incentive model, it is important to have funding incentives (an industrial logic), but we also need to emphasize the intrinsic value of building infrastructure, such as new and interesting research methods and insights.
- Scale and time are major barriers for developing infrastructure, especially with the constant disruptions in available technology. It becomes nearly impossible to create a unified,

interoperable system with so many variables changing rapidly.

- Data infrastructure has also taken different models depending on the domain/ community implementing the system. Whether the infrastructure is more like connections to “islands” of data resources or a way to share key data sets continually, what does data infrastructure look like for the digital humanities community?

Presentation of RDA Infrastructure

Larry Lannom began the discussion of specific RDA data sharing infrastructure which may be valuable to the DH community. He emphasized that the early outputs are about precision and data management. The first working groups include Data Foundations and Terminology, Data Type Registries, PID Information Types, Practical Policies, and the Metadata Standards Directory. For a specific example of how RDA deliverables are produced and adopted, Larry discussed the Data Type Registries WG. The problem this working group tackled is a lack of knowledge about a data sets' format and type as it is parsed and shared beyond the original context. As a solution, DTR created a “registry framework allowing easy registration and identification of precise data definitions at multiple levels of granularity that can be reference from within data sets and/or data set metadata.” The example provided was of a stream gauge measurement data type that specifies the particular and complicated units that are collected from a stream gauge instrument. The use case for this infrastructure is to allow a user to download or acquire a data set. If the type is unknown, the user can query the registry to resolve its type definitions, relationships, and other metadata. Ideally, in the future, this information can also be used to find appropriate services and providers. The DTR is currently in prototyping with four projects, and other users are considering adopting it. Larry concluded by mentioning the international, bi-annual working group collaboration meetings for co-chairs (an opportunity for a potential DH working group to meet and collaborate with other working groups), and the data fabric interest group (a space for making the larger connections and infrastructure components of the data lifecycle.

Kim Fortun began with her experiences working as the co-chair of the RDA interest group Digital Practices in History and Ethnography. Kim introduced several complications for adapting STEM developed data sharing infrastructure into a DH project. For one, the humanities have been classically averse to using digital tools because there is a reigning fear that digital techniques destroy the data's context and that digital tools do not allow the scholar to think through a theory of language. As a corollary to this, many STEM fields distrust the 'subjective' nature of humanistic inquiry. However, the subjective criticism is misplaced: the individual research of humanists is complex and

grounded in a “sedimented genealogy of thought and tradition”. This makes adopting infrastructure a fraught yet necessary process. Kim also shared a layer metaphor for understanding how data is generated and shared in the humanities; the empirical humanities create data through the selection and curation process.

Discussion of Use Cases in relation to RDA Infrastructure

Bridget Almas provided a synthesis of the use cases submitted by meeting participants prior to the meeting. There are currently many barriers to data sharing and collaboration: it is complex, there is limited funding (with a great deal of competition by potential collaborators), a lack of incentives, and enabling factors. Any infrastructure(s) must be constantly evolving and adapting to changing parts. A great deal already exists, such as repository software, database and PID systems, workflow and visualizations tools, and metadata management and image services, etc. To create the connections between these services, tools, and systems, the digital humanities need “glue” in the form of unambiguous data types and supporting services, consistent application of standards, policies and governance strategies, novel tools and services for rapidly changing research environments, and clear guidance on how to implement new infrastructure to DH use cases. The incentives for creating data and data sharing infrastructure in the humanities is increased funding, but it also includes the ability to reach a wider audience, scholarship sustained beyond a single project, and a better use of resources (without the need to redevelop tools for every new project). Funding incentives should challenge and require projects to either produce interoperable and sustainable shared data tools or to reuse existing and proven solutions.

Some of the unique challenges for humanities infrastructure are:

- complex and diverse data citation practices for dynamic, static, and linked data,
- a mixture of both qualitative and quantitative data types, with uncertainty and a need for attribution to contributors,
- complex access issues and rights in the form of privacy and copyrighted materials,
- new publication models, with the generation of short pieces, open access material, and new annotations,
- scholars/ data produces are also changing to include crowd sourced material, student researchers, and machine generated knowledge,
- there is hybridization and fusion between traditional disciplinary boundaries, and

- any new tools must be generalized and sustainable across many fields while simultaneously supporting domain-specific functions.

The RDA can provide support for these efforts with specific working groups (the Data Type Registries and PID collections), a global forum for communication and expert input, and action-focused groups that avoid being paralyzed by indecision. The question for consideration during the workshop is: Can RDA provide a space for existing projects to "unlock" their data silos and develop solutions that will be adopted across disciplines while solving the original project's urgent needs?

Reactions, Comments, Etc. from Workshop Participants

The following are ideas or comments from participants during the discussion session:

- There may be a problem with collaborating in the humanities across national boundaries because the humanities, unlike the natural sciences, are frequently invested in nation building or localized projects (examples: history and language studies). We must find unifying issues or attack the grand challenges in the humanities in order to overcome this division rooted in nationalism/ localization. The RDA is also an international organization that can help overcome national politics around projects and funding.
    - An example of a "grand challenge" for DH is ensuring linked data is interoperable.
    - Another example of a DH grand challenge is the plurality of languages that data appears in. The data is not all in English, which means it may be difficult to adopt existing solutions.

- The working group structure may need to be inverted for the Digital Humanities. For example, RDA currently proposes a piece of infrastructure to develop. In DH, however, there are already existing pieces of data sharing infrastructure. Is it feasible to create a working group dedicated to generalizing (or customizing for a new setting) these existing solutions and using the working group to amplify its use? The RDA/US representatives seemed receptive to this idea that generalizing an existing product is a meaningful output with one important caveat:
    - RDA is technology neutral, so it will not certify any outputs as "the one" solution to adopt, nor will the RDA endorse a specific, already existing platform. RDA has maintained this mission by emphasizing interoperability for all products produced.

- Another potential DH working group could evaluate RDA's existing (largely STEM produced) outputs for their applicability in a humanities setting. This would encourage broader adoption and provide “push back” where these solutions are lacking.

- Brett Bobley (NEH) asked the workshop to articulate the value added from building DH infrastructure within RDA. Several responses were:
  - RDA means that projects are not “reinventing the wheel” every time they begin a new initiative.
  - RDA also provides standing to challenge entrenched disciplinary inertia; for example: selecting a metadata standard to adopt can be an excruciatingly slow process. With RDA's stamp of approval that a metadata standard works, it is easier to implement a data solution.
  - RDA exposes individuals to new, interdisciplinary perspectives and the data sharing problems each domain has. These problems and solutions can help DH generate new ideas through exposure.
  - RDA's international network means that there are more potential collaborators and more community feedback.
  - RDA's domain groups and broader plenary sessions means that conversations exist in both specific detail and interdisciplinary spaces.
  - The working group structure means that any outputs have the backing of a community, not just an individual project, and that cultural capital can help with amplification.
  - RDA is also practical and work focused, meaning there is an eventual end to simply talking about problems without ever taking action.
  - The RDA outputs also provide authority and guidance on which system, tool, or policy to use.
  - RDA also provides a larger context and scale that most DH projects currently cannot access. A durable community of practice is important for infrastructure development to work.
  - RDA can be a way to prove to funders that a project is “playing well with others” and is deserving of funding.

- Scale for digital humanities projects has some similarities with scientific domains, yet it is also distinct in several ways. One, the data sets can be much smaller, but the data is varied, complex, and requires parametrics and metadata as part of the dataset. (An example provided was for an archeology project which used traditional, material objects as data while also generating drone scans and 3D models of the ancient site.) One participant mentioned that there is a belief in

computer science that the humanities are “data poor”, that is, they cannot work at scale. However, this is untrue: the humanities have a great deal of data, they simply require the tools and techniques to extract the data.

- Dissemination of research/ scholarship is also an issue for the digital humanities. It is difficult to declare something “published” if it is “living” and has ongoing annotations. How does one site a humanities data set that is on-going? And at what level of detail?
    - One participant suggested “corpus enhancement” be considered a form of publishing: e.g. adding temporal and geospatial tags to existing narratives so they can be queried by place and time.
    - Publishing visualizations is also a form of interpretation and can be considered scholarship.
    - A difficulty with publishing data for the humanities is ensuring that it has the necessary meta-data AND actionable machine tags so that the data can continue to be part of the research process.

- Another challenge is co-publishing qualitative data with quantitative data from other domains. There is a need for infrastructure that supports diverse data types. For example: it is difficult for an epidemiologist to compare the narrative portion of a hospital intake form to the numeric metrics provided because most software tools do not offer an analytic layer for the free-form answers.

- Preservation may be the issue which unites digital humanities with other domains because at the level of infrastructural issues, the needs and use cases are the same.

RDA/US Response

Sayeed asked the RDA/US representatives present why the first set of working group deliverables have not been adopted broadly?

In response, Beth Plale (co-chair of RDA's Technical Advisory Board) introduced a diffusion model that has important stages for the “adoption pipeline”. First, RDA must build awareness of outputs, then support interested individuals seeking more information. Next, these individuals must evaluate the infrastructure for fit with their own project, and finally RDA must facilitate

implementation and testing in their environment. Fran Berman reassured the workshop that RDA/US has many people at various stages of this pipeline, and resources are currently dedicated to recruiting more adopters. Larry Lannom concluded by saying that if every project and working group succeeded, then the organization has been too conservative. RDA must invest time and resources in novel approaches that might end with failure, but the community has still learned something from these experiences. RDA would love to see more humanists and social scientists join the community. Fran Berman emphasized that RDA is a place to facilitate the development of socio-technical infrastructure.

Funders Comments

Before concluding the workshop, Sayeed asked the funders to comment on their impression of the discussion and any suggestions as a funding representative.

Trevor Owens emphasized the importance of libraries as the traditional home for humanities data and archival work. These must continue to be connected to DH projects as a hub for services, networking, and resource sharing. His organization also supports training for librarians and works with their constituencies to make sure online archive platforms last. The next round of proposals for IMLS will be in September 2015.

Helen Cullyer (Mellon) discussed several provocations she noticed. One of the major challenges for the future of DH is material under copyright or data that is born digital. There will also be a need to create/ adapt to a broader publishing ecosystem. She also noted that it is important that new infrastructure programs and initiatives are not top-down. Helen sees useful priorities for DH (and Mellon in particular) as: linked open data, annotation tools, restricted access data sharing, interoperability of image data tools, new presses for DH research, and funding for data curation fellowships. Mellon's funding process begins by sending a 2 page letter of interest summarizing the proposed project.

During the discussion, the following issue was described: thousands of faculty members around the world do not associate the humanities and data or recognize that DH offers new research methods; even a large portion of humanists do not associate their scholarship and research with generating or using data. This means that there is a major communication barrier and long road ahead for introducing the Research Data Alliance to the broader community: need a value proposition for why RDA is helpful for the field. DH scholars implicitly think that the RDA is not for them, it is a “science thing.” NEH is

distinctive in that it funds new methods, so there is a space for new proposals to ask for a grant line for RDA participation that will support the development of new infrastructure and lend credibility to the broader impact of grant proposals. The predominate infrastructure need is also human—DH needs training institutes and people who act as "glue" to share and implement the technical solutions.

Next Steps and Action Items

1. Encourage workshop participants to join RDA and investigate working and interest groups for "good fit" with their projects.

2. Host a second workshop or outreach meeting to further refine the issues raised in this meeting. One suggestion is to host a *Birds of a Feather* meeting for the Digital Humanities at the coming RDA 6th Plenary in Paris. Any co-chairs or co-organizers are supported for travel.

3. RDA/US will continue to provide support for meetings and outreach to engage with the digital humanities community. In June, Bridget Almas will travel to DH2015 (the international meeting sponsored by the Alliance for Digital Humanities Organizations) to discuss and introduce the Research Data Alliance at a special lunch event. Brett and Greg are also traveling to Canada in June for a Digital Humanities meeting and may be able to describe RDA as relevant while there.

4. The RDA/US office will provide flyers for the DH2015 meeting, and these press materials will be made available on the RDA website for other DH scholars to utilize at other professional meetings and conferences.