# “Good Things Come in Small Packets”: How (Inter)national Digital Research Infrastructure can support “Small Data” Humanities and Cultural Heritage research

Daniel Paul O’Donnell, University of Lethbridge, for the “Good Things” Research Team

The purpose of this whitepaper is to describe a largely unrecognised and unsupported but very common research data management (RDM) use-case: that of the traditional “Small Data” Humanities and Cultural Heritage (HCH) research project producing or working with “primary source” research data (e.g. digital facsimiles, recordings, and representations of cultural objects/activities). This paper complements the submission from the Canadian Society for Digital Humanities/Société canadienne des humanités numériques (CSDH/SCHN), which is concerned with the case of research and data in the Digital Humanities more broadly, including in such small data contexts.

As we shall argue in this paper, the kind of traditional data and RDM use-case we are discussing here has gone largely unrecognised by Digital Research Infrastructure (DRI) developers and policymakers — in part because the nature, size, methods of production, and purpose of these datasets are quite different from data production and management in other disciplines, and in part because the data themselves are not always understood *as* data (or their management as an RDM problem) by the relevant research community (e.g. 1,2).

The result is that large quantities of small-project HCH research data are poorly managed and maintained and that often extremely well curated datasets produced by HCH researchers remain

- invisible, siloed, or difficult to access by Big Data researchers (where such access is appropriate and ethical);
- unnecessarily expensive to produce and maintain; and, as a result,
- in danger of premature loss or obsolescence to researchers and the wider community.

The specific tools and techniques required to address these problems already mostly exist within the international DRI ecosystem. However, we are aware of no single system or provider that includes them all. We are also aware of no system or provider that specifically supports the workflow, use-case, and datatypes we describe here, including such Humanities-focussed projects as Humanities Commons. The “Good Things come in Small Packages” Partnership was created in order to promote recognition and support of this RDM use-case and has been in discussions with several developers, providers, and policymakers interested in promoting and supporting improved RDM practices for traditional “Small Data” HCH research. This exercise represents an excellent opportunity for Canada, through NDRIO, to lead internationally in supporting this societally important research.

Because this use-case is so different from those of other disciplines, this paper begins with a description of the traditional “small-data” HCH research project as an RDM problem before briefly answering the CFP questions.

## Small, Thick, and Slow: Research data in the traditional humanities

The first thing to realise is the degree to which traditional HCH research data and use-cases differ from those of Science, Technology, Engineering, and Medicine (STEM). Where STEM typically involves datasets produced through experiment or observation (i.e. “capta” or “taken things”), “Small Data” HCH often involves the deep analysis and intensive curation of very small data sets acquired to act as “primary sources” — that is to say as facsimiles, records, or models of cultural objects, events, or activities (i.e. as “data” or “given things”) that are intended to be used by other researchers for further analysis (for a summary see 3).

The Good Things partnership describes these kinds of data as “Small,” “Thick,” and “Slow”:

- **Small** in the sense that they can often involve the representation of a single real-world object (the reproduction of a single novel, painting, recording, etc.), a small collection of related objects (e.g. a *fonds*

or some other small collection gathered on thematic, historical, functional, or other bases), or multiple representations of the same object (e.g. photographic facsimile, transcription, and editorial text);

- **Thick** in the sense that they are usually intensely and richly curated and contextualised (i.e. thickly described), with thousands (even tens or hundreds of thousands) of words of what is in essence contextual metadata — definitions, historical analysis, comparisons to related data(sets), citations, etc.;
- **Slow** in the sense that the definition and analysis of individual data points can continue for generations — and, as Borgman argues, often represent "particularly prized acts of scholarship in their own right" (4).[1]

The focus on detailed and careful contextualisation of very small data sets over what is often several generations means that HCH research data themselves are often of extremely high quality in terms of the care with which they are produced and the extensiveness and quality of what is, again, in essence, their accompanying metadata. Unfortunately, however, HCH researchers have no real tradition of understanding such data *as* data (and indeed at times argue against understanding them this way, see 1,2). Because data-contextualisation rather than data-production or -collection is the primary act of scholarship, there is no tradition of management, publication, or even archiving of such representational data as stand-alone research outputs (libraries and archives are in this context better considered as a type of infrastructure than research projects). In the digital age, this tradition has persisted in the largely "craft" model that digital HCH data projects have adopted for the design and construction of their datasets (5,6) — a model in which every project is independently responsible for the design, encoding, storage, and publication of its own data and analysis (and indeed, almost always, interface). As a rule, this means that the data themselves can usually only be accessed indirectly — via the larger project, using a bespoke organisational and identification system — rather than directly via standardised Digital Data Repositories (DDRs) and Persistent Identifiers (PIDs). It also leaves these data vulnerable to early obsolescence or data loss — as code breaks, links rot, and funding provided for research rather than maintenance dries up.

An example of just what is lost can be seen in the case of medieval manuscript photography. Since the first modern digital editions were published in the 1990s, editors of European medieval texts have almost invariably included full-colour, high-definition photographs of the manuscripts upon which they are basing their editions (7,8). Just as invariably, these photographs are accompanied by a huge apparatus of what can be understood as thickly described metadata and references to other objects: diplomatic transcriptions and "reading" texts, collations establishing correspondences among different copies of the same text, codicological and paleographic descriptions, bibliographic histories, cultural and literary commentary (9,10). While best practice has long encouraged researchers to use descriptive file names (11,12), and while recent projects such as the International Image Interoperability Framework (IIIF; 13) have developed Application Programming Interfaces (APIs) for delivery and referencing, most photographs in most editions were published outside of these frameworks and standards — in custom locations, using custom naming conventions, and, usually, without an explicit machine-readable association between data and metadata. This means, in turn, that almost 30 years' worth of expert-curated and -described cultural data is, for all intents and purposes, invisible to machine-driven "Big-Data" research — and in danger of early obsolescence and loss. The raw material is there; but lack of common standards prevents discovery for what would otherwise be a large virtual collection of intensely curated data and metadata.

The fact that nearly every digital project dealing with Western European Medieval Manuscripts has designed what is in effect a custom digital data infrastructure — its own way of organising and displaying data and metadata, its own identifiers and locations, and its own way of linking analysis to the underlying data — invariably increases costs and decreases access (14,15). The fact that similar practices are ubiquitous across the Digital and traditional Humanities means that almost all digital cultural data is much more expensive to produce and curate, and far more

[1] In this context, it is worth noting that traditional HCH research data is almost always non-rivalrous: it is the process of analysis and contextualisation rather than discovery that is the most-prized act of scholarship and individual disciplines can return productively to the same few well-known data points over and over again for years (e.g. Jane Austen Studies).

vulnerable to loss and obsolescence, than should be tolerable. This is perhaps particularly true in the case of projects dealing with Indigenous, culturally sensitive, or endangered data — projects that face similar technological hurdles as other HCH projects with regard to the management and publication of their primary material, but also have additional obligations — such as respecting First Nations ownership and authorship (16,17), ensuring the right to create value from Indigenous data in ways that are grounded in Indigenous worldviews, and understanding the historical context and power differentials involved in their collection (18): a project that goes dark because it has an unsustainable architecture is a project that has failed in its obligation to ensure that data remains in or gives back to the community from which it came in an accessible format.

## Digital Research Infrastructure as a path forward

One solution to this problem is the creation of a DRI that has been built and is operated with an understanding of this prototypical HCH research data use-case. Such an infrastructure would see itself as part of a research-to-publication workflow and help HCH researchers and research projects — and their publishers and users — to understand how careful and sustainable RDM practices can reduce their production and maintenance costs and improve longevity, discoverability, access, and impact (Figure 1 shows a conceptual model developed in partnership with Zenodo). It would provide incentives to researchers to deposit their data in DDRs by ensuring that such deposits could be subsequently and easily incorporated in the kind of traditional contextualising research project — edition, database, study, or application — that remains the primary and most valued research outputs in these disciplines.

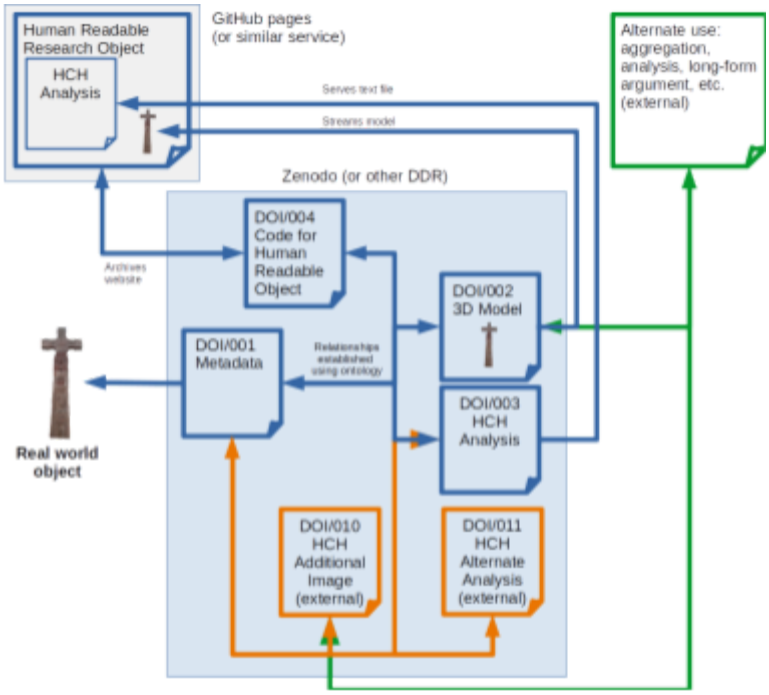


Figure 1: Conceptual workflow showing DRI-based "Small-Data" publication workflow

Currently, because such contextualised publication is the primary and most valued output for HCH research, adding data to a stand-alone DDR requires HCH researchers to, in essence, deposit their data twice: once within their custom-designed application using their own protocols and identifiers and then a second time in the DDR using the DDR's protocols and standards. A better approach would ensure that DDRs — and other aspects of a centrally maintained DRI — were both incorporated into a suitable workflow *and* made this workflow easier to use than the traditional, custom-built alternative. The basic method — understanding a DRI as an infrastructure upon which other research tools operate, using methods such as PIDs to store, access, and transport data — is relatively established in STEM (see 19,20). But we are unaware of any attempts to develop a similarly atomic system for the publication of HCH data — the far more significant problem in those domains.

The problem is not, for the most part technical: most of the major features required are already found in one or another of the major DRI projects — e.g. DDRs that can store data deposits as individual FAIR objects and serve out files to external sites (Figshare); APIs that allow URLs to be built on the basis of PIDs (Zenodo); support for the development and publishing of contextualising scholarly outputs (GitHub); robust Linked Open Data (LOD) capabilities (Zenodo); the ability to version and archive websites (GitHub + Zenodo). Likewise features that are not yet implemented — such as the ability to use arbitrary or discipline-specific ontologies for typing semantic links — are either in partial development (e.g. 21–23) or would not appear to be particularly difficult to add.

The real problem is leadership, resources, and a holistic commitment both to understand the nature of the RDM problems presented by traditional HCH research project data and to work with leading HCH researchers, research organisations, libraries, archives, and publishers to ensure that this crucial use-case is "baked in" to DRIs from the beginning. As a country that punches far above its weight in Digital Humanities research, Canada and the NDRIO are well positioned to ensure that even the small, thick and slow get the support they need.

## Current Issues

- What are the main DRI tools, services and/or resources you currently use in your research?
    - *Data repositories, PIDs, hosting and archiving services*.
- Do you have access to all the DRI tools, services and/or resources you need for your research? What are they? What is missing?
    - *While most of the services that our research requires are present in one form or another somewhere, no comprehensive system has been established with the needs of a traditional "Small Data" HCH project in mind. A Zenodo-like deposit system would also be good.*
- What are your biggest challenges accessing and using the DRI tools, services and/or resources that do exist and are available to you?
    - *Lack of awareness on the part of HCH researchers, and a failure on the part of DRI developers to understand and accommodate the needs of HCH researchers, publishers, etc.*

## Future DRI State

- What is your vision for a cohesive Canadian DRI ecosystem that would fulfill your research needs?
    - *The vision promoted here is for a system that understands itself as part of an HCH research-to-publication workflow — in which data-infrastructure is a seamless part of a PID-based publication workflow.*
- What are the types of DRI tools, services and/or resources you would like to use, or envision using, in the future?
    - *Streaming, API-access to data stored in DDRs — the precise mix is less important than the understanding that a DRI should support the publication of HCH research data in context.*
    - *What is crucial, however, is training and community work to make the change from current practices to an Open, FAIR (and, as appropriate, CARE)-complaint DRI ecosystem.*
- What challenges do you foresee while using integrated DRI tools, services and/or resources?
    - *By its nature HCH research problematizes solutions: it is ultimately impossible to develop or impose a single solution for HCH publication. But there is still much core work that can be done to ensure that traditionally-focussed HCH projects are optimised for long term sustainability.*

## How to Bridge the Gap

- What are the tools, services and/or resources NDRIO should leverage to achieve your desired future state?
    - *Rather than proposing specific tools, we'd argue that NDRIO should consider an environment scan — looking at specific use-cases with discipline experts and comparing how such use-cases are or are not supported by existing DRIs. It is striking the degree to which most of the tools, services, and resources already exist, but are distributed unevenly across different systems.*
- How do you see NDRIO's role in addressing current gaps in the national DRI ecosystem?
    - *As a national agency in a country where there a) appears to be reasonably good integration and collaboration among the three main agencies; b) is an appetite for interdisciplinary work (e.g. Frontiers); and c) funding exists for the whole research cycle from initial idea (the various discovery and development grants) to publication (ASJ), we believe that NDRIO is well-situated internationally to take the interdisciplinary approach proposed here.*

## Works cited


1. Marche S. Literature is not Data: Against Digital Humanities. Los Angeles Review of Books [Internet]. 2012 [cited 2013 Mar 30]; Available from: http://www.lareviewofbooks.org/article.php?id=1040&fulltext=1
2. Fish S. Mind Your P's and B's: The Digital Humanities and Interpretation [Internet]. Opinionator. 2012 [cited 2013 Mar 30]. Available from: https://bit.ly/3gKM3mC
3. O'Donnell DP. The bird in hand: Humanities research in the age of open data. In: Figshare, editor. The state of open data: A selection of analyses and articles about open data, curated by Figshare. London: Digital Science; 2016. p. 34–5. (Digital Science Report).
4. Borgman CL. Scholarship in the digital age: information, infrastructure, and the Internet. Cambridge, Mass: MIT Press; 2007. 336 p.
5. Flanders J. The Productive Unease of 21st-century Digital Scholarship. 2009 [cited 2013 May 12];3(3). Available from: http://digitalhumanities.org/dhq/vol/3/3/000055/000055.html
6. Jockers M, Flanders J. A Matter of Scale. Faculty Publications -- Department of English [Internet]. 2013 Mar 18; Available from: https://digitalcommons.unl.edu/englishfacpubs/106
7. O'Donnell DP. Move over: Learning to read and write with novel technology. Scholarly and Research Communication [Internet]. 2012;3(4). Available from: http://www.src-online.ca/index.php/src/article/view/68
8. O'Donnell DP. Back to the future: what digital editors can learn from print editorial practice. Literary and linguistic computing [Internet]. 2009; Available from: https://academic.oup.com/dsh/article-abstract/24/1/113/947909
9. McGillivray M, Asgar-Deen T, editors. Geoffrey Chaucer's Book of the Duchess a hypertext edition. University of Calgary Press; 1999.
10. Kiernan KS. Electronic Beowulf. London, Ann Arbor: British Library. University of Michigan Press; 1999.
11. Lee SD, O'Donnell DP. From manuscript to computer. In: Owen-Crocker GR, Cesario M, editors. Working with Anglo-saxon manuscripts. University of Exeter Press; 2009.
12. Terras MM. Digital images for the information professional. Aldershot, England; Burlington, VT: Ashgate; 2008.
13. IIIF. Home [Internet]. International Image Interoperability Framework. Available from: https://iiif.io/
14. O'Donnell DP. The Doomsday Machine, or, "If you build it, will they still come ten years from now?": What Medievalists working in digital media can do to ensure the longevity of their research. Heroic Age [Internet]. 2004;7. Available from: http://www.mun.ca/mst/heroicage/issues/7/odonnell.html
15. Copland C, Carrell S, Davidson G, Grandfield V, O'Donnell DP. Kiernan, Kevin S. 2015. Electronic Beowulf-http://ebeowulf. uky. edu/ebeo4. 0/CD/main. html. Digital Medievalist. 2016;10.
16. Schnarch B. Ownership, Control, Access, and Possession (OCAP) or Self-Determination Applied to Research: A Critical Analysis of Contemporary First Nations Research and Some Options for First Nations Communities. JAH. 2004;1(1):80–95.
17. Bliss H, Genee I, Junker M-O, O'Donnell DP. "Credit where credit is due": Authorship and Attribution in Algonquian Language Digital Resources. IDEAH [Internet]. 2020 [cited 2020 Jun 1]; Available from: https://ideah.pubpub.org/pub/z4ime60s/release/1
18. CARE Principles of Indigenous Data Governance [Internet]. Global Indigenous Data Alliance. [cited 2020 Nov 12]. Available from: https://www.gida-global.org/care
19. Bosman J, Kramer B. 101 Innovations in Scholarly Communication [Internet]. 101 Innovations in Scholarly Communication. Available from: https://innoscholcomm.silk.co/
20. Kramer B, Bosman J. Innovations in scholarly communication - global survey on research tool usage. F1000Research. 2016 Apr 18;5:692.
21. Seltmann KC, Pénzes Z, Yoder MJ, Bertone MA, Deans AR. Utilizing Descriptive Statements from the Biodiversity Heritage Library to Expand the Hymenoptera Anatomy Ontology. Moreau CS, editor. PLoS ONE. 2013 Feb 18;8(2):e55674.
22. Biodiversity Literature Repository [Internet]. Zenodo; 2013. Available from: https://zenodo.org/record/3475439
23. Cui H, Jiang K (Yang), Sanyal PP. From text to RDF triple store: An application for biodiversity literature. Proc Am Soc Info Sci Tech. 2010 Nov 1;47(1):1–2.