December 2017

Rethinking the Potential of Documentation of Culture as a Data Gathering Practice

Tomasz Umerle
Institute of Literary Research of the Polish Academy of Sciences, tomasz.umerle@ibl.waw.pl

In this article, I will be examining the *documentation of culture* (DoC). This is the practice of gathering and producing the information and data relevant to intangible cultural phenomena. DoC is generally practiced by teams of documentalists generally outside GLAM institutions (i.e., galleries, libraries, archives and museums). As such, it differs from the preservation and description of concrete objects of cultural heritage done within those institutions.[1]

I will be talking about DoC projects—like bibliographies, filmographies and dictionaries in the form of books or databases—which are also independent from the academic library and information science (LIS) community and infrastructure. As such, they realize their potential as digital data-centric initiatives later, and in a different cultural context, than those of GLAM institutions. I will be using examples of Polish DoC projects to address the specific problems that DoC projects and services are facing today.

For decades, DoC projects in Poland have been institutionally independent from library systems and information science departments and have been a part of research institutes, universities, or cultural institutions. As such, they have developed fundamental documentation resources which have allowed DoC to fulfill the role of being an auxiliary discipline to the humanities.

Today DoC projects in Poland deal with issues that undermine their traditional position. DoC practitioners increasingly find it difficult to clearly define and communicate their role: a) in relation to LIS; b) to the broader academic community; and c) as a practice in the digital age of information overload. In this article, I will briefly outline the specificity of DoC, its main issues and possible solutions to its current problems.

What Does DoC Document?

The examples of long-term Polish documentation initiatives include institutes within the Polish Academy of Sciences (e.g., Institute of Literary Research, Institute of History, Institute of Art), higher education departments (The Polish National Film, Television and Theater School), cultural public organizations (Theater Institute, Polish Library of Songs[2]), and cultural non-governmental organizations (Polish Music Information Centre, Bibliography of Polish Ethnography).

For decades, teams of documentalists have worked within these institutions on specific projects important for all those interested in culture in Poland. Here are a few examples of such projects:

[1] It is worth noting that many of the issues discussed in this article apply also to the practices of librarians, museum cataloguers, etc., but looking outside the context of preserving material resources offers an interesting viewpoint on issues specific to DoC projects.

[2] Polish Library of Songs is a creator of an online database called Digital Library of Songs, which gathers metadata of Polish songs and to some extent grants access to audio versions of songs, but it does not primarily store music albums, CDs, publications, etc.

- Polish Film Database: http://www.filmpolski.pl/fp/index.php
- Encyclopedia of Theater (multiple databases): http://www.encyklopediateatru.pl/
- Polish Literary Bibliography: http://pbl.ibl.poznan.pl
- Polish Biographical Dictionary: http://www.psb.pan.krakow.pl/index.php/en/
- Digital Library of Polish Song: http://www.bibliotekapiosenki.pl/glowna
- Polish classical music database: http://www.polmic.pl/index.php?lang=pl
- Bibliography of Polish Etnography: http://odie.ptl.info.pl/
- Non-digitized examples, such as Polish Artistic Life (calendar), Biographical Dictionary of Polish Theater, Dictionary of Polish Artists

Documentation projects in Poland are mostly connected to either scientific or cultural institutions and serve the purposes of registering certain domains of culture and/or humanities, and supporting research. Some of these initiatives are also institutionally connected to archival or librarian practices, nonetheless their documentarian projects are dedicated to registering certain cultural phenomena, not just the material objects they preserve or archive. That is why we might look at them as *data gathering practices*.

The idea behind such documentation projects may be construed as such:

DoC Project (with a given methodology)	**Mediator (evidence)**: Documents	**Object of documentation**: Living culture

Which translates into specific examples, such as:

Polish Literary Bibliography	**Mediator (evidence)**: Books, journals	**Object of documentation**: What is considered "literary"
Theater Institute database of theatrical repertoire	**Mediator (evidence)**: Theatrical programs, other collateral	**Object of documentation**: All the theatrical works ever staged in Poland

It is debatable what types of documents should be included in these kinds of documentation projects, and there are always scientific and methodological considerations; they serve a mediating role in the more general task of documenting cultural phenomena. The question that DoC documentalists ask themselves is, "Are school literary competitions, or amateur theater, or Facebook posts, etc., important enough as a cultural phenomenon for us to

DOI: 10.35492/docam/4/2/15

document them?" rather than, "Should a specific, physically identified set of documents be processed?"

To fully understand this attitude towards documentation practices, it is best to once again evoke the famous Suzanne Briet's example of the antelope, a living animal which—according to Briet—becomes a document when it is catalogued for the first time: "*L'antilope cataloguée est un document initial et les autres documents sont des documents seconds ou dérivés*" (Briet, 1951, p. 8). What contemporary researchers find fascinating and groundbreaking in Briet's work is the fact that in Briet's categories a documentalist is in fact preoccupied with a certain state of a "thing" or "being" that becomes a document when it is framed as such by society. It makes the definition of *document* fluid, and definitely not restricted to set of traditional documentation resources (books, journals, etc.).

DoC projects free from the duty to document specific resources stored in their archives or magazines face the "antelope dilemma" on a regular basis. What happens when a poem/song is written/captured and registered, then published on the Internet? Is this a part of a literary/musical culture that should be registered? As a documentalist working on creating a database I am not worried about archiving this document; it would take me half a minute to introduce a new record into my database concerning this specific phenomenon. In the end, I am definitely facing what Briet would call "*la poeme/chanson cataloguée.*" Of course, I will not do that right away, due to strict methodological rules of my DoC project, but those rules are subject to discussion. And this discussion is being performed within the framework envisioned by Briet. Going beyond this, the question of DoC is not, "What new kinds of catalogues can document these captured antelopes?" but rather, "What kinds of antelopes are running wild in the world?"

According to Briet, as Buckland suggests (Buckland, 1997, 1998), this makes the document physical evidence of an object we are really interested in: "*Un document est une preuve à l'appui d'un fait*" (Briet, 1951, p. 7). And in the digital age, it is easier than ever before to gather enormous amounts of evidence of artistic practices which in turn make DoC—unexpectedly—heavily influenced by the contemporary data-centric culture on the one hand, and on the other it engages documentalists in continuous theoretical discussions about the nature of art, music, performance, etc., which Buckland also incisively notes:

> This situation is reminiscent of discussions of how an image is made art by framing it as art. Did Briet mean that just as "art" is made art by "framing" (i.e. treating) it as art, so an object becomes a "document" when it is treated as a document, i.e. as a physical or symbolic sign, preserved or recorded, intended to represent, to reconstruct, or to demonstrate a physical or conceptual phenomenon? (Buckland, 1997, p. 806)

What is unexpected in DoC becoming a part of data-centric culture is, firstly, the fact that DoC projects have mostly originated from traditional documentation resources. The above-mentioned DoC resources—dictionaries, bibliographies, filmographies, databases—are the effects of documentalists' work of reading, listening to and watching culture firsthand and producing knowledge for a certain subdiscipline in structured, or semi-structured forms (e.g., handwritten, printed, electronic databases). Secondly, DoC projects have not been a part of the LIS infrastructure (neither library software, nor modern, usually costly, scientific databases), so the projects themselves are methodologically more diverse than typical LIS projects and frequently use custom metadata schemas. While LIS institutions have worked together on standardization of metadata and compatibility of their software, DoC projects methodologies were not subject to regulations or discussions that were supposed to prepare data-gathering initiatives to function properly in the digital age. We might add that the DoC tradition of metadata creation is connected in a complex way to well-known other traditions, namely library and data management approaches (Burnett, Ng & Park, 1999). It is freed from the duty to serve users interested in a specific stored object, but it has not from the start been connected to computer-assisted data production.

What DoC projects share with the basic ideology of the data-centric culture is the Briet-like open and dynamic definition of documentation. Documentation for DoC is a practice of documenting living culture evidenced by socially constructed documents and focused on creating and providing information on culture, not preserving actual cultural objects, which theoretically means that DoC projects are ideologically well-equipped to introduce digital tools and standards into their practices. At the same time this basic ideological commonality faces serious technological challenges, which makes the position of DoC in contemporary data- and documentation-driven culture complex and interesting.

Unexpected Consequences of Data-Centric Culture

DoC in Poland has gathered vast and unique resources of what we could call today *cultural or humanistic data*, created in accordance with the best academic standards of documentation. The primary documentary resource of DoC projects was cultural and humanistic information that was traditionally treated as an auxiliary resource for the humanities that materialized in the form of books routinely distributed as such. Digitization steadily reshapes the identity of DoC projects into projects that are almost entirely data-centric, making documentation specialists creators, curators and managers of datasets.[3]

A rough estimate tells us that if data gathered through DoC projects were

[3] From this point of view DoC projects also possess their own concrete documents—like GLAM institutions—although they do not possess cultural objects (primary cultural documents like books, movies, screenplays of plays, sculptures etc.).

DOI: 10.35492/docam/4/2/15

digitized, the potential database would consist of at least 10 million entries.[4] Only a small part, approximately 20 percent, of these resources has been digitized—some of them have not even been published (e.g., handwritten or printed card catalogues). Nonetheless it is an obvious goal of each institution to digitize as much knowledge as they can in the form of structured data, which is treated as a universal "necessity of our time." For Lund, this necessity is embedded in the state of contemporary culture and the dominance of data in information science:

> Necessity emerges from information overload in modernity and from the needs of scientists and engineers (as well as others) to have quick and efficient access to the newest information in their own and neighboring subject domains. For Briet, specialized cultures make up the culture of modernity, but the culture of modernity is overall driven by greater needs for "efficiency" and "dynamism," and this includes in intellectual domains. (Day, 2014, p. 9–10)

It seems as though DoC is, in a sense, in a perfect spot—it is focused on creating (meta)data that is a privileged form of science communication nowadays, and it can participate in meaningful discussions about the current state of documents in the age of renewed interest in documentation studies and general instability of the definition of *document*. However, digitization funding in Poland (and not only there) is mostly dedicated to cultural objects (which are mostly in possession of libraries and archives), not their metadata, which can be produced by entities different from those that digitize and/or possess certain object. And in the process of digitization that is being routinely funded, there is a general reluctance to treat metadata creation as a necessary component of digitization projects, which is a universal issue noted by researchers:

> [There is a] continuing aversion to spending some time and money on creating metadata and the belief that there may be some technical panacea to remove this tiresome burden… There may be a number of reasons for this but one that stands out for us is that in institutional terms the role of librarians and information professionals has been sidelined and downgraded in favour of IT services since the inception of the worldwide-web. Typically, IT departments have little expertise in the area of information management—but believe they do. Metadata creation, cataloguing and classification may not be currently fashionable but it does provide an essential means to find materials and avoid "digital oblivion." (Casey, Proven & Dripps, 2006)

[4] Data gathered via questionnaire from the following institutions: Institute of History, Institute of Art, and Institute of Literary Research of the Polish Academy of Sciences; The Polish National Film, Television and Theater School; Theater Institute; Polish Library of Songs; Polish Music Information Centre; Bibliography of Polish Ethnography.

The notion of IT utopia is not the only reason for this reluctance. In the end, the vision of content queries replacing metadata-based queries is still far from realization and far from becoming a popular belief, especially in the realm of culture or humanities, which are susceptible to terminological and historical instability. This reluctance should also be attributed to the psychological consequences of an information overload culture that creates a popular assumption that there is "already enough data" so easily accessible that it seems counterintuitive that the creation of data may be a reasonable time-, people- and money-consuming enterprise (Hoq, 2014, pp. 55–57).

The other reason for the focus on digitization of objects is connected to the users' perception of digitized cultural heritage. As Lund (2009) proposes, there are three aspects of documents—physical, social and mental. Lund encourages analyzing the ways in which they interact with each other. The digitization of cultural objects focuses on creating a strong mental experience of physicality—the pictures, galleries, photos and gifs are the core elements of digital libraries, and they serve as the evidence of the ability of digitization to make the user *feel* as if they experienced the actual, physical object.

From this point of view, the digitization of cultural objects is a more concrete practice as compared to the creation and digitization of metadata, which result in more abstract practices of introducing records into a database or an entry in a dictionary (or even enriching existing records).

Making Metadata Worthwhile

DoC projects are and will be metadata-centric (and the obvious work of linking metadata to actual objects is just a part of enriching records/entries). And as such they have to address the above-mentioned reluctance to support metadata creation by embracing the identity of metadata creators or providers, which means rethinking DoC projects in terms of treating them as sources of cultural and research datasets.

DoC projects in Poland have gathered knowledge on Polish culture, which is divided into cultural domains (theater, music, literature, art, history, ethnology, etc.). However, one of the main and obvious advantages of data-based research should be the ability of researchers to freely manipulate data from different domains of culture, which in turn is possible only when handlers of data put enough effort into assuring that their discipline-specific data is stored in a manner that facilitates interdisciplinary research. And this task might be achieved only through the multi-institutional effort to strengthen cooperation between DoC institutions.

Assuring the interoperability and interconnectedness of metadata might be obvious from the vantage point of LIS, due to its infrastructural embeddedness in standardized metadata formats (despite their fluid nature), but DoC projects are more methodologically diverse. And it is not just a Polish specificity. Australian Humanities Networked Infrastructure (HuNI), an

DOI: 10.35492/docam/4/2/15

endeavor created by Australian DoC institutions, is a data aggregate that is built largely of custom metadata schemas. HuNI creators tried to incorporate these datasets into existing ontologies, but resigned due to the very nature of metadata concerning culture and humanities gathered by DoC institutions:

> Significant technical and conceptual difficulties became clear with this approach. In part, these challenges arise from broad disciplinary shifts within the humanities, away from a traditional focus on measuring the value and meaning of cultural artefacts to recognizing the import of cultural flows and the dynamic nature of cultural infrastructure (itself understood as a creative process and catalyst of social and environmental amenity). (Verhoeven & Burrows, 2014)

In this context, one of the differences between DoC and LIS metadata—we might infer, having in mind the analogies between DoC and Briet's definition of documents—that arises from DoC's commitment to document living culture is a more flexible and discipline-specific metadata. The interconnectedness and interoperability of metadata has not traditionally been a priority of disciplinary DoC projects; however, they have been prioritized within LIS in the past decades.

The second issue with building datasets that facilitate interdisciplinary research is the fact that still the majority of DoC resources have not been fully digitized, due to the above-mentioned reluctance to fund metadata creation (or, in this case, to be precise, *remediation*). Dictionaries, calendars or bibliographies need to be transformed into structured forms of data storage (e.g., relational databases), which is a task that most likely demands using innovative natural language processing and machine learning technologies together with human expertise.

Cooperation between DoC projects in creating interdisciplinary and open access research datasets faces two mid-term and long-term challenges: 1) building ontologies for custom data schemas; and 2) obtaining funding for innovative digitization projects.

For the time being, what DoC institutions can focus on is convincing the academic community, and other users of scientific information, of the obvious advantages of building interdisciplinary datasets. And this can be done through engaging building smaller data collections dedicated to interdisciplinary domains of culture that do not require implementing universal solutions.

Designing interdisciplinary machines

The long-term goal of DoC might be seen as a form of adaptation to "modern information and communication technologies [that] are, above all, *devices and systems that function through both sociocultural and technical logics*" (Day, 2014, p. 63). It may serve, for Day, as the testament to how documentation sometimes plays the role of a follower of contemporary domination of data **"**as

privileged, and indeed sometimes, as the only form, of knowledge and experience and so as the governing mediator for understanding and actions" (Day, 2014, p. 151). What documentation sees as a more efficient way to explore and communicate knowledge is a form of sociopolitical power that has colonized our everyday existence. The data-centric culture creates an environment in which people cease to control documentation tools and themselves become the objects of documentation for powerful devices such as search engines, browsers and other applications.

Indeed, Day's (2014) *Indexing It All* and other literature critical (in general or in specific) of documentation projects that embrace data science ideas or tools should serve as a constant benchmark for the implementation of these ideas and tools:

> The opacity of what machines are doing in the background of our lives creates illusions of fairness, equity, privacy, or full access to everything. None of these is assured or necessarily true for all seekers and users of information. The growing issue of automated control over scholarship requires that we become better informed about the nature of computation and its control over what is accessible from and delivered to our devices. We should push for, if not demand, transparency, education, and dialogue about the implications of the choices the machines make (through human programming choices) and their potential risks, principles and standards, and mitigation strategies and options—plus far more than we can elucidate at this time. (Alexander et al., 2016, p. 13)

DoC projects should infer from this critical literature three main ideas, while embracing data science concepts. Firstly, the duty to first and foremost document a certain domain of culture treated as cultural heritage. The scope of documentation should not be influenced only by the current interests of users (because the concept of cultural heritage always demands a long-term perspective), especially if it leads to its narrowing (focusing on documenting "most read" or "most searchable" items, etc.). What should be constantly influenced by users' current needs is the design of tools and devices introduced by DoC projects. Secondly, DoC project should implement technologies that guarantee preserving rich and diverse methodologies of various DoC projects as primary (like semantic web or linked data) or secondary (flexible noSQL databases) tools for making interdisciplinary datasets connected. Thirdly, DoC projects should be transparent as to how digital tools, even those that serve just to present knowledge, might influence users' image of culture.

References

Alexander, B., Barrett, K., Cumming, S., Herron, P., Holland, C., Keane, K., Ogburn, J., Orlowitz, J., Thomas, M. A., & Tsao, J. (2016). Report from the Information Overload and Underload Workgroup, Open

DOI: 10.35492/docam/4/2/15

Scholarship Initiative Proceedings, Volume 1. Retrieved from https://journals.gmu.edu/osi/article/view/1383/1204

Briet, S. (1951). *Qu'est-ce que la documentation?* Paris: Éditions Documentaires, Industrielles et Techniques.

Buckland, M. 1997. What Is a "document"? *Journal of the American Society for Information Science, 48*(9), 804–809.

Buckland, M. 1998. What is a "digital document"? *Document Numérique, 2*(2), 221–230.

Burnett, K., Ng, K. B. & Park, S. (1999). A comparison of the two traditions of metadata development. *Journal of the Association for Information Science and Technology, 50*(13), 1209–1217.

Casey, J., Proven, J., & Dripps, D. (2006). Geronimo's Cadillac: Lessons for learning object repositories. Retrieved from http://www.csfic.ecs.soton.ac.uk/Casey.doc

Hoq, K. M. G. (2014). Information overload: Causes, consequences and remedies—A study. *Philosophy and Progress, 55–56*(1–2), 50–68.

Lund, N. W. (2009). Document theory. *Annual Review of Information Science and Technology, 43*, 1–55.

Verhoeven, D., & Burrows, T. (2014). Digital Scholarship in the humanities and creative arts: The HuNI virtual laboratory. *EDUCAUSE Review*. Retrieved from https://er.educause.edu/articles/2014/6/digital-scholarship-in-the-humanities-and-creative-arts-the-huni-virtual-laboratory

Day, R. E. (2014). *Indexing it all: The subject in the age of documentation, information, and data*. Cambridge, MA: The MIT Press.