ACCESS QUALITY METRICS FOR NET ART

Xiao Ma
Independent Researcher
USA
xm75@cornell.edu
0000-0001-6134-3531

Dragan Espenschied
Rhizome
USA/Germany
dragan.espenschied@rhizome.org
0000-0003-1968-6172

Lyndsey Jane Moulds
Rhizome
USA
lyndsey.moulds@rhizome.org
0000-0002-4858-0417

Rhizome's ArtBase is a public archive holding copies of more than 800 works of net art. Most pieces allow for different points of access: they might be available live from a Rhizome web server or alternatively from a web archive, with both versions potentially incomplete and in different states of restoration. Visitors might view these works via a period-adequate browser in an emulator, or whatever setup they are running on their devices. Discussed below is a system using technical metadata and curatorial information to calculate an access quality score that can help visitors choose which artworks, versions, and modes of access will best meet their needs.

Keywords – Digital Art, Net Art, Access, Emulation
Conference Topics – Innovation; Resilience

I. PRESENTING NET ART IN AN ARCHIVAL CONTEXT

Rhizome's ArtBase, an online archive started in 1999, holds pieces of net art that have entered the collection through different mechanisms, including open accession (1999-2008), curation (2011-2020), and open calls (starting 2021) [1]. Methods used to package and stabilize the works varied depending on what tools and concepts were available at the time (artist-submitted file copies, web archiving, disk imaging, etc.) as well as how the artworks were made and conceptualized as objects [2].

The quality of access to archived born-digital art can be thought of as a result of two factors: first, the availability of stable and complete resources, and second, the capabilities of the software environment used to perform the works. Net art introduces further complexity: many works are not self-contained, but instead present a "blurry" object boundary [3] that may not be easily understood or accurately demarcated in the moment of archival. As an example, an artist might submit an incomplete set of files to the archive, and omissions might not become apparent until external resources fall offline much later.

Additionally, net art is usually produced for and accessed via whatever devices and software internet users have available and is in most cases not tied to a canonical software environment. Over time, this mix of operating systems, browsers, and other applications to access online materials change in their forms and capabilities. These changes range from the drastic, like deprecation of certain file formats and programming languages, to less noticeable changes such as the deprecation of features allowing browsers to open popup windows, play MIDI music, or draw certain UI widgets [4].

Preservation and restoration actions can in many cases retroactively supply missing resources and, via emulation, prepare software environments that provide the best possible circumstances for the digital artifacts to be performed. The result of each preservation action is a new "variant" of the artwork [5]. Each variant is composed of a set of stabilized artifacts and a software and network environment.

For each of these variants, Rhizome aims to provide an access quality score that is an expression of these possible states of a variant. This is done to direct newcomers to highlights of the collection and manage users' expectations of artworks that expose deficiencies. The score is especially useful while an artwork is transitioning from being best accessible on the live web to being best represented in a controlled, encapsulated environment constructed for preservation purposes. Users will have to make the tradeoff between accessing a variant of the artwork that is integrated into the present landscape of the internet but may be less functional, versus a variant that is more separate from the live internet but offers a reliable, reproducible performance. The access quality score can guide them to the variant that fits their intention for access.

18th International Conference on Digital Preservation
iPres 2022, Glasgow, Scotland.

Described below is the data model and process required to compute a single access quality value per variant that can be displayed as a simple 3 level "stop light" indicator on access links: green variants should be expected to be as complete as possible and have all current preservation goals met, yellow variants have known problems, red variants must be expected to be incomplete or at least partly non-functional [6].

II. Data Model and Data Sources

ArtBase is built as a Linked Open Data repository with Wikibase. Artworks are modeled in the following manner:

- Variants are represented as a combination of *artifacts* and *machines*.
- Artifacts can be collections of files, disk and media images, containers, web archives, etc. [7], with their components described based on the PRONOM file format registry.
- Machines are configured virtual machines, emulators, and containers managed via EaaS, or an approximation of the software environments widely used (see below). They are described by the *software* that is installed on the disk image they boot from.
- Each software is a self-contained, installed package with its *capabilities* described by the data formats it can handle, again using the PRONOM file format registry.
- Finally, the capabilities of a machine represent the sum of the capabilities of the software installed on it, which then can be matched against the components of the artifacts.

This technical data can be automatically generated (artifact composition can be determined by a tool like Siegfried) and observed and recorded in experiments (supported data types can be elicited by trying to run software with specimens of that data type).

One special type of machine is the "default access machine," representing the capabilities of an assumed contemporary software environment that approximates the lowest common denominator of different devices, operating systems, browsers, etc. that are available to regular web users. New machines are described in sync with the general landscape of contemporary software changing, and assigned to variants that are accessed "directly," rather than via an emulator. In addition to representing an approximation of currently available capabilities, modeling a projected default access machine can be used to project the effects of upcoming software changes on a collection, for instance when a browser vendor announces that support for a particular video codec or plugin will be discontinued.

Table 1 Data Model

subject	predicate	object	note
machine	has part	software	Software installed on machine
software	has part	software	Optional nesting for bundles
software	handles	data format	Capabilities of a software package
artifact	made of	data format	Artifact has at least one occurrence of a data format
variant	has artifact	artifact	Artifact used in variant
variant	has machine	machine	Machine used this variant
variant	handles	data format	Optional curatorial information over-riding machine values
variant	made of	data format	Optional curatorial information over-riding machine values
	relevance	relevance value	Qualifier holding a multiplicator value indicating a data format's relevance for the intended purpose of the variant.
variant	access quality	access quality value	Computed value expressing the variant's access quality.

Information recording the capabilities of software based on PRONOM has already been proven meaningful to create matches of existing configured machines with artifacts in a library context [8][9]. When applied in the context of art and access quality, the considerations need to be slightly different, in sometimes counterintuitive ways:

1) There is no correlation between the number of occurrences of a certain data format (as in "how many files of this type are part of an artifact?") with the relevance of that data format for an artwork's performance. For instance, a Microsoft Word file being part of an artifact might be an artist's

18th International Conference on Digital Preservation
iPres 2022, Glasgow, Scotland.

description of their work submitted as a package to Rhizome and not be referenced or linked to in the actual artwork at all. As a result, the machine used for access does not need to provide software to render this file if the goal is to present the artwork. In another access context, like the analysis or exhibition of artists' descriptions of their work, the capability to render this type of file would be essential. Each access scenario needs to be modeled as its own variant, combining the same artifact with different machines. A machine needs to provide the capabilities to render a data type if it occurs more than zero times and is relevant. The machine does not need to provide capabilities if the data type occurs zero times or is deemed irrelevant. The actual number of occurrences greater than one is not producing better scoring results. Even if only one jpeg file out of a set of ten is deemed relevant, the machine used will have to support that format.

2) Unidentifiable or misidentified data formats are common in digital art, in which oftentimes artists employ tools that have little relevance in the library field and hence are not represented in the PRONOM registry, or at least not at the required level of detail that would enable correct automatic detection. Additionally, the adherence to standards like "well-formed XML" that would make format detection more reliable has little relevance in the production of digital art. 34% of the works in ArtBase contain at least one occurrence of an unidentified data format. "Clean" solutions—implementing a new format detection rule, or manually assigning a synthetic format ID to every occurrence—is considered too laborious and demanding too much expert knowledge to implement in day-to-day collection management. Instead, both cases are handled via a value manually added to the variant that denotes the relevance of a particular data format for the access quality calculation. Since the relevance of a file format is tied to the intention of making the variant available, it does not make sense to record it with the artifact.

III. From Data to Readiness Score

Defining Readiness Score: Baseline

We have established above that a variant is not just the static files (artifacts) associated with it but is a combination of artifacts performed in a particular environment (machine).

$$Variant = Artifact \times Machine$$

Therefore, we define a "readiness" score for a variant as a feature of the variant that indicates how likely the variant's performance can be perceived as complete by the user.

The most basic definition of a readiness score can be:

$$readiness\_score\_0_{artifacts,machine} = \frac{num\_supported\_data\_formats}{num\_data\_formats}$$

Here is a toy example:

Table II Example Variants

Variant ID	Artifact	Artifact composition	Machine ID	Supported (inferred)
1	1	doc	97	True
1	1	jpeg	97	True
1	1	mp3	97	True
2	1	doc	98	True
2	1	jpeg	98	False
2	1	mp3	98	False

Let's say we have a variant with ID 1 composed of an artifact that contains three types of files, doc, jpeg, and mp3. We have two machines that might support the artifacts, therefore we have two variant IDs. The readiness score of variant 1 is 1 because all file types are supported. The readiness score of variant 2 is 0.33 because only one file type is supported.

1. Variant-Specific Relevance Score

The baseline score assumes that each data type is equally important to the artwork. As established above, depending on the purpose of the variant being made accessible, some data types might be crucial for the intended performance while others do not require support.

To make the readiness score more accurate, we can augment the baseline with human curatorial information. A human curator can examine each variant and assign a relevance score to file types on a Likert scale of 1-5, with 1 being not important at all (the viewer's experience will just be fine if this file type is not supported), and 5 being very important (the experience is meaningless without this file type being supported).

Essentially, for each variant, we can use the importance score as a weight to modify baseline readiness score.

In this example, the human curator will examine the variant 1, and assign scores of importance based on the artifact and machine combination.

18th International Conference on Digital Preservation
iPres 2022, Glasgow, Scotland.

Table III Variants with Curatorial Relevance Rating

Variant ID	Artifact ID	Artifact composition	Machine ID	Supported (inferred)		Relevance score (curatorial label)
1	1	doc	97	True		1
1	1	jpeg	97	True		1
1	1	mp3	97	True		5
2	1	doc	98	True		5
2	1	jpeg	98	False		1
2	1	mp3	98	False		1

1. Ignore the unknown files in the readiness score computation: this method is easy but may not be accurate.

2. Default "null" to false or true for consistency: easy to implement but may not be accurate.

3. Have human curators examine the unknown file, correct the file type if known, and assign a supported true/false label, as well as an importance score to override the unknown. This method is more label intensive, yet considered to be attainable in day-to-day collection care, and should provide the most accurate data and estimate of readiness score. (See Table IV.)

Table IV Variants with Curatorial Relevance and Support Rating

Variant ID	Artifact ID	Artifact composition	Machine ID	Supported (inferred)	Supported (curatorial label)	Relevance score (curatorial label)
1	1	doc	97	True		1
1	1	jpeg	97	True		1
1	1	mp3	97	True		5
1	1	*unknown*	97	*Null*	*False*	5
2	1	doc	98	True		5
2	1	jpeg	98	False		1
2	1	mp3	98	False		1
2	1	*unknown*	98	*Null*	*True*	3

The curatorially supported readiness score can be calculated as follows:

$$readiness\_score\_1_{artifacts,machine} = \frac{\Sigma\, int(supported\_by\_machine) * relevance\_score}{\Sigma\, relevance\_score}$$

In the toy example above, variant 1 would have a readiness score 1 of (1+1+5) / (1+1+5) = 1 (very good support) again; while variant 2 would have a readiness score 1 of (1*5) / (5+1+1) = 0.7 (good support), reflecting that the human curator has indicated that on the machine ID 98 environment, jpeg files are not relevant enough to require support. The augmented readiness score therefore incorporates curatorial knowledge and can be a more accurate estimate than the baseline readiness score.

2. Handling Unknown Data Types

As established above, a consistent way to handle unknown file types in the computing of the readiness score is essential for the digital art use-case.

In the case that file types are unknown, the automated pipeline to infer whether a file is supported would return null value. In these cases, we have a few options:

3. Computing and Presentation

Once a variant's readiness score is computed it can be stored as a property of that variant. On the user interface, the value can be used to draw the access quality stoplight indicator and for providing ranking and filtering functions.

Each time an element involved in the computation of this value changes—such as a new data format being detected in an artifact due to a PRONOM update, a software installed on a machine is found out to have different capabilities than originally thought, etc.—the value must be re-computed.

IV. FUTURE WORK

There is, of course, room to improve in terms of accuracy for the readiness score as defined above.

The most important omission in the current calculation is concerning the grade of completeness of available artifacts, which we plan to include in an upcoming version.

In addition, we plan to leverage timestamps as a source of improving data quality. Say if an artwork was created in the year 2000, and the machine contained software released earlier or much later,

18th International Conference on Digital Preservation
iPres 2022, Glasgow, Scotland.

we can infer that the machine might not support a file type even if the corresponding identifiers would match.

Finally, we can explore machine learning techniques to learn to parse the composition of artworks based on curatorial information. For example, we can try to predict the importance score of a file type—based on features such as file size, last modified time, and past curatorial importance labels. These computing techniques can further improve the efficiency of digital preservation staff and more quickly provide users with an access quality score.

References

[1] Rossenova, L. (2020) "ArtBase Archive—Context and History: Discovery Phase and User Research 2017–2019". Available from: https://lozanaross.github.io/phd-portfolio/docs/1_Report_ARTBASE-HISTORY_2020.pdf

[2] Espenschied, D. 2021. "Digital Objecthood," in Selçuk Artut, Osman Serhat Karaman, Cemal Yılmaz, eds. Technological Arts Preservation, Sabancı University, Istanbul, 2021. ISBN 978-625-7329-16-3 https://www.sakipsabancimuzesi.org/en/page/technological-arts-preservation

[3] Espenschied, D., Rechert, K. "Fencing Apparently Infinite Objects," in: Proceedings of iPRES 2018. DOI 10.17605/OSF.IO/6F2NM. https://phaidra.univie.ac.at/view/o:923620

[4] Espenschied, D., Kreymer, I. "Oldweb.today: Browsing the Past Web with Browsers from the Past" in Gomes, D., Demidova, E., Winters, J., Risse, Th. (Eds.), The Past Web, Springer, 2021. ISBN 978-3-030-63290-8

[5] Rossenova, L., de Wild, K., and Espenschied, D. "Provenance for Internet Art: Using the W3C PROV data model," in: Proceedings of the 16th International Conference on Digital Preservation iPRES 2019, Amsterdam, The Netherlands, pp.297-305. https://osf.io/qc9u5/

[6] Rossenova, L (2020) "ArtBase Redesign—Prototype Development: Design Exploration and Evaluation 2018–2019". Available from: https://lozanaross.github.io/phd-portfolio/docs/4_Report_DESIGN_EXPLORATION_2020.pdf

[7] Espenschied, D. "Artifacts and Infrastructure." Rhizome, 2021. https://almanac.rhizome.org/pages/artifacts-infrastructure

[8] Giessl, J., Gieschke, R., Rechert, K. and Cochrane, E. "Automating the Selection of Emulated Rendering Environments for Born-Digital Data-Sets," in: Proceedings of the 25th International Conference on Theory and Practice of Digital Libraries, TPDL 2021, Virtual Event, September 13–17, 2021. Springer. https://link.springer.com/chapter/10.1007/978-3-030-86324-1_12

[9] Thornton, K. "Wikidata for Digital Preservationists. DPC Technology Watch Guidance Note". 2021. Digital Preservation Coalition. DOI 10.7207/twgn21-19

18th International Conference on Digital Preservation
iPres 2022, Glasgow, Scotland.