Soft Materials

Exploring Data Quality in Data-Integrated Design

Maya Przybylski
University of Waterloo, Canada

Introduction – Soft Materials

The increasing availability of real-world data is having a profound effect on the practice of architecture. The integration of data-driven computational methods into design processes has transformed the way architects work across all project phases – from design and planning to execution and operation.¹

This integration presents new modes of engagement for architects and designers in allied fields whereby they play the role of producers of digital technologies – developing custom computational processes that are embedded as agents in the design phases and beyond. This practice, labelled here as software embedded design (SED), supports wide-ranging work: supporting undertakings such as form generation, automated analysis, advanced visualizations, and interaction and responsiveness. Often, these custom

computational processes remain in the studio; the code is active while the project is in development, and when complete, its output is disconnected from the computational apparatus when deployed in the physical world. Yet, increasingly these processes go beyond the pre-deployment phases to include scenarios where the components remain online and continue to actively mediate project outcomes and behaviors through time as space – post deployment, in the real-world. By embedding software as an active element, a soft material, charged with mediating function and performance over time, designers are effectively adding computational components to their project’s material assemblies.

As software is embedded in this way, a new set of responsibilities emerges for its producer – in this case, the architect. The increasing links between computational elements, material elements and project performance and behaviors adds complexity to how we assess project impacts and outcomes. We can consider a broad range of issues in this regard including things like technical performance, spatial effect and programmatic support. However, as part of the material assembly, we argue that the data and algorithms defining soft materials need to be understood also as playing a direct role in shaping the socio-cultural outcomes of the design work.

This argument is supported by key offerings from the field of Software Studies which position software – its actual lines of code and data, not just its effects – as a material practice with both social and spatial outcomes. Software Studies seeks to crack open the prevalent representations of software as a black box and to understand its inner workings with respect to instructing technologies how to act, and work done in this arena has successfully demonstrated that software is a social-material production with profound influence on everyday life.

These perspectives support a reconceptualization of ethical literacy in software embedded design; by including soft materials, such as data and algorithms, in our conception of a project’s material assembly they implicitly become a part of the matrix defining the works’ ethical dimensions. Thus, an expansion of knowledge and methods to support a more complete engagement with computational components embedded in such work is necessary. As producers of the material assemblages constituting our design work, we need to be literate and possess agency with respect to the social, cultural
and political effects across the entire assembly – and this includes not only the material outside the computer, but the material inside the machine, and the connection between the two. Specifically, what are the necessary methodologies to support these a socio-cultural understanding of the softer materials of design work? How do we understand software’s agency as a spatial agent? How should we shape designers’ relationship with custom software to ensure its spatial agency is acknowledged, managed and leveraged towards the most ethical outcomes?

This is particularly important as a maturing SED field is on the cusp of moving from test-spaces with relatively minimal risks, to real-world applications with potentially more serious outcomes. This transition demands that designers assess their capacity to engage fully with adopted extra-disciplinary techniques. As SED projects move to consider seriously their social and cultural consequences, designers need to exercise the same criticality they have shown the physical project components towards the virtual project components operating inside the computer and the relationship between the two. The data we integrate and the algorithms we code need to be understood and examined as designed project elements with agency over project outcomes – not only from a technical perspective, but also with respect to social and cultural impacts.

Specifically, designers need to integrate questions and criticality around data sources, data format and selection, algorithm design and coding culture in order to confront the spatial and sociocultural agency of their data-driven works. As we enter new frontiers in computational design, it is no longer enough to only have methods that integrate data in design projects but new methods are needed that engage with it critically throughout the design process and beyond.

Data Quality & Terrain Nominal as Key Concepts to Better Understand Soft Materials

This paper focuses on the discovery and articulation of methods aimed at expanding the degree to which designers relate specifically to the data components of soft materials (algorithms and coding culture will be tackled in future work). It presents the key concept of data quality (DQ) as described...
in disciplines with more established histories of valuing data-literacy, such as information science, geography and other social sciences, and goes on to examine the potential for DQ assessment in SED work.

Data Quality: A Multi-dimensional Problem

Thinking about data in terms of its quality offers an entry point for forming a critical assessment of data consumed and produced in SED work. The details of such a process, such as first identifying the characteristics that should be evaluated and then carrying out the assessment, is an active research front rooted in the information sciences but also present in other disciplines such as geography, sociology and computer science. DQ is often presented as a multi-dimensional concept emphasizing a varying set of characteristics depending on its assessors’ perspective. Thus, fitness-for-use is widely adopted as a core principle in DQ implying that data quality is relative: its assessment depends on its use and data appropriate for one use may not be appropriate for another.

Numerous researchers across various fields have developed DQ frameworks which identify and group various data attributes which constitute holistic pictures of ‘quality’. Knight and Burn offer an extensive examination of the twelve most widely accepted DQ frameworks. They review the literature and enumerate the DQ dimensions listed in the studied frameworks and account

<table>
<thead>
<tr>
<th>DIMENSION</th>
<th>DEFINITIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>extent to which data are correct, reliable and certified free of error</td>
</tr>
<tr>
<td>Consistency</td>
<td>extent to which information is presented in the same format and compatible with previous data</td>
</tr>
<tr>
<td>Security</td>
<td>extent to which access to information is restricted appropriately to maintain its security</td>
</tr>
<tr>
<td>Timeliness</td>
<td>extent to which the information is sufficiently up-to-date for the task at hand</td>
</tr>
<tr>
<td>Completeness</td>
<td>extent to which information is not missing and is of sufficient breadth and depth for the task at hand</td>
</tr>
<tr>
<td>Concise</td>
<td>extent to which information is compactly represented without being overwhelming</td>
</tr>
<tr>
<td>Reliability</td>
<td>extent to which information is correct and reliable</td>
</tr>
<tr>
<td>Understandability</td>
<td>extent to which data are clear without ambiguity and easily comprehended</td>
</tr>
<tr>
<td>Accessibility</td>
<td>extent to which information is available, or easily and quickly retrievable</td>
</tr>
<tr>
<td>Objectivity</td>
<td>extent to which information is unbiased, unprejudiced and impartial</td>
</tr>
<tr>
<td>Relevancy / Usefulness</td>
<td>extent to which information is applicable and helpful for the task at hand</td>
</tr>
<tr>
<td>Usability</td>
<td>extent to which information is clear and easily used</td>
</tr>
<tr>
<td>Amount of Data</td>
<td>extent to which the quantity or volume of available data is appropriate</td>
</tr>
<tr>
<td>Believability</td>
<td>extent to which information is regarded as true and credible</td>
</tr>
<tr>
<td>Navigation</td>
<td>extent to which data are easily found and linked to</td>
</tr>
<tr>
<td>Reputation</td>
<td>extent to which information is highly regarded in terms of source or content</td>
</tr>
<tr>
<td>Efficiency</td>
<td>extent to which data are able to meet the information needs for the task at hand</td>
</tr>
<tr>
<td>Value-Added</td>
<td>extent to which information is beneficial, provides advantage from its use</td>
</tr>
</tbody>
</table>


for how often they appear (Image 2). This list is useful as it exposes a preliminary longlist of concerns to be incorporated in forming critical relationships with data consumption.

Of the twelve presented by Knight and Burn, the framework created by Wang and Strong (WS) entitled “A Conceptual Framework for Data Quality” is selected here for closer examination. The WS framework is refined to address four key categories (Image 2). 1.) **Intrinsic DQ** addresses the extent to which data values are in conformance with actual or true values. This category includes **accuracy**, the extent to which data are correct, but also includes **objectivity**, the extent to which information is unbiased, unprejudiced and impartial. 2.) **Context DQ** deals with the extent to which data are applicable to the task. Key components of this category include **value-added**, the extent to which information is beneficial, provides advantages from its use, and **relevancy**, the extent to which information is applicable and helpful for the task at hand. 3.) **Representational DQ** addresses the extent to which data are presented in an intelligible and clear manner. 4.) **Accessibility DQ** concentrates on the extent to which data are available and obtainable.

We posit that the WS framework is well suited to serve as a starting point in developing a preliminary framework for DQ in SED. First, the WS framework departs from previous approaches by focusing on DQ from the perspective of the data consumer, not the producer. Further, the WS framework’s top-level categorization aligns well with familiar design considerations such as performance, form, context, and representation. Finally, the WS framework has successfully served as a foundation for more refined frameworks.
developed for specific domains. The WS framework offers a structured guide to organizing dimensions against which designers can form critical relationships with data. The framework’s core concerns are familiar enough to designers to offer a familiar point of entry yet broad enough to support the integration of SED specific concerns.

Terrain Nominal

The conceptual organization and terms provided by DQ frameworks facilitate a more critical engagement with data integrated in design projects. By exposing the multidimensional quality of data, frameworks identify some of the questions designers need to ask of the data being consumed in their projects.

In order to satisfy a given dimension of a DQ framework, the data must be understood by the extent to which it fulfills the specific dimension. This terminology reminds us that DQ is relative: the extent or degree to which a dimension needs to be met must be specified by the designer. It is only when the designer specifies the degree to which the data needs to be accurate (or relevant, objective, timely, concise…) that DQ assessment can take place. The sum of these specifications defines the terrain nominal which declares, implicitly and explicitly, the required level of abstraction and generalization relative to the real-world phenomena the data attempts to capture. Terrain nominal, accepts that there is no objective reality or essential truth the data must capture and instead focuses on establishing a specification corresponding to a perspective of the phenomena which is tuned to the task at hand. In some cases, data is known to be subjective, incomplete, or inexact; by accepting ‘shortfalls’ such as these in the conceptual model the data may still be fit-for-use and meaningful to the project. Thus, to judge fitness-for-use, the designer needs to not only assess the data against the specification, but also consider the effectiveness and limitations of the conceptual model itself. (Image 3).

Terrain Nominal is also a direct reminder of the inherent subjectivity captured in software systems. Software can be understood as “a form of subjectivity – the software constructs sensoriums, that each piece of software constructs ways of seeing, knowing and doing in the world that at once contain a model of that part of the world it ostensibly pertains to and that also shape it every time it is used.”

15 Veregin, 1999
16 Veregin, 1999
17 Veregin, 1999
Towards a Preliminary DQ Framework for Data-Integrated Design

In order to make some of these thoughts more concrete, we use an existing SED project and carry out a post-mortem reflection on how its DQ assessment may transpire. Specifically, an initial mapping of these extra disciplinary understandings of DQ into the design domain is tested through the proposal of a preliminary framework for assessing data quality in a project entitled OnTheLine (project credits omitted for blind review).

OnTheLine was conceived as an exhibition which converts a transit line into a gallery connecting transit riders with the rich collection of destinations located along the route. The project team created a suite of physical installations and digital interfaces aimed at shaping and presenting a collective identity of the region. Utilizing bus shelters, buses, a centralized project display, an interactive website, and a mobile app as media for information transmission, OnTheLine both distributes and collects information about activities, events, and destinations along the transit corridor (Image 4).

Two complementary sets of data contribute to the unified presentation of destinations: first, the project presents a curated and centralized set of local cultural destinations and activities that has been pre-assembled.
by the project team from a suite of local information resources; second, an emergent set of destinations submitted by the public, via social media mechanisms such as twitter hashtags, augments the list of presented destinations (Image 5). This hybrid data set constitutes a growing directory of local information, engaging a diverse range of participants as project collaborators. The project’s various sites, its collection and use of data, and its supporting media and technologies, are combined into a comprehensive framework for enabling the city as a site of participation, thereby offering a mechanism by which to sense the city and enable its strategic development and management.

**Image 5** OnTheLine: Conceived as an exhibition which converts a transit line into a gallery connecting transit riders with the rich collection of destinations located along the route, the project offers a comprehensive framework for enabling the city as a site of participation, thereby offering a mechanism by which to sense the city.

**Image 6** The OnTheLine’s Datascape: two primary data streams, curated and emergent, are combined into a suite of physical installations and digital interfaces.
Taking lessons from the DQ assessment context established above the following steps are outlined for conducting a high-level assessment of the DQ used in *OnTheLine*:

1. Establish *OnTheLine*'s socio-cultural and technical objectives in order to establish the terrain-nominal.
2. Identify the DQ dimensions to be assessed.
3. Prioritize selected DQ dimensions.
4. Develop specific assessment techniques for prioritized dimensions.

The literature identifies three approaches to studying DQ: intuitive, theoretical, and empirical. The *intuitive* approach is taken when the selection of DQ attributes is based on the assessor’s experience or intuitive understanding about which characteristics are “important”. While this approach has some shortfalls – the assessment is limited to the DQ knowledge existing within a specific project by not integrating DQ concerns held by external parties – it does allow for a quicker identification of relevant DQ attributes. By being relatively easy to formulate, the intuitive approach also allows for each study to select its own set of attributes most relevant to its particular goals.

The intuitive approach has been selected to develop a preliminary DQ framework for *OnTheLine*. An exhaustive DQ assessment of the project is beyond the scope of this paper. Instead a subset of the project’s data is strategically culled to highlight the diverse set of questions and analytical outcomes across the several datasets used in the project.

Results & Discussion

Image 6 presents a summary of Steps 1 – 3 outlined above. A sampling of specifications driving the key technical and socio-cultural objectives of the project is offered (left) to establish a conceptual model of the phenomena that the data is meant to capture – namely the name and location of notable sites along the transit line. These specifications establish the key considerations informing the prioritization of the Wang & Strong DQ attributes (center). Connecting lines are drawn to communicate which specification relates to which DQ attribute. A Preliminary intuitive ranking of attributes is also provided.
Step 4 demands specific assessment techniques to be developed. During these early stages of this research, Radial Coordinate Maps are used to create multivariate representations of the data’s quality in an accessible way\textsuperscript{21}. A total of three sets of Radial Maps are produced. The first set, presented in Image 6, establishes target values for DQ Attributes based on the projects stated objectives. The two subsequent sets of maps are produced as part of Step 4. Image 7 presents the differential between Curated Data and target values according to the DQ framework’s three main categories. The final set of maps (Image 8) examines the quality of Accuracy as a single attribute in more detail.

Image 7 presents a high-level intuitive assessment of OnTheLine’s Curated Data, assembled by the project team from a variety of community-level sources, such as the Ontario Heritage Trust and Waterloo Regional Tourism group, and subjected to manual processing, verification and selection. Processing steps, such as cross-referencing municipal addresses with geocoding results, increased ratings for accuracy and believability. The timeliness attributes received a low rating as some data sources were dated over one year prior to the project implementation and thus increased the probability of venue being closed or otherwise out-of-date. Objectivity was also scored relatively low: the selection process behind the data’s inclusion (immediate availability was a key driver and less available sources were excluded) did little to ensure an unbiased and impartial dataset.

Image 8 presents a more detailed comparison of the Accuracy attribute across the Curated (CData) and Emergent (EData) data sets used in the OnTheLine. Here, due to the data’s geospatial nature, accuracy is unpacked to include three additional dimensions by which to better understand its quality: temporal accuracy, spatial accuracy and thematic accuracy\textsuperscript{22}. Independent examination of each of these sub-properties exposes discrepancies between the two datasets. On the one hand, the real-time features of the EData overcame some of the timeliness issues of the CData discussed above. On the other hand, EData was more susceptible to lower levels of relevance and higher probabilities for mis-categorization, resulting in a lower rating for thematic accuracy. Further, users’ smartphone location settings caused the precision of spatial information to vary widely across EData entries. Meanwhile, the CData was subject to spatial accuracy verification which ensured highly precise spatial data.


\textsuperscript{22} Veregin, 1999
Conclusion & Future Research

The preliminary work presented in this paper begins to formulate an avenue for building richer engagement with data in SED work. Perhaps most clear in this work is the fact that confronting data quality is a multi-dimensional challenge demanding the articulation of new questions and methods for designers.

This work valued quick intuitive assessment carried out within a guiding framework. This technique is useful in that it demonstrates an accessible approach for designers to begin to formulate some degree of criticality around the data they use in their work. Future work will integrate more robust assessment techniques. Specifically, recent developments in surveying design spaces using Performance Maps in order to capture multivariate properties\(^{23}\) will be examined as a way to move beyond the Radial Coordinate Maps used here.

At the framework level itself, more work needs to be done to meaningfully assemble DQ attributes especially relevant to SED. The adopted frameworks were effective starting points, but they are biased towards the concerns of external disciplines. Reviewing already existing SED work...
in order to build an inventory of possible SED-focused attributes is one approach. Surveying SED authors in order to understand existing data-related objectives and values is another. This will also illuminate questions around the way designers conceive and evaluate the terrain nominal which drives the representation of the captured phenomenon. Specific challenges around incorporating qualitative data quality concerns, such as objectivity, also demand further examination.

Of course, data is only one aspect of Soft Materials. Similar avenues should be established and supported in order to ask similar questions of the algorithms used in SED work. In general, the inherent subjectivity of software calls on us to be explicit in the assumptions and biases embedded by us in the models used in the software in our projects. These biases may arise for numerous reasons: 1) technical limitations – coders’ ability is limited so aggressive decisions around how to represent the problem are made, 2) resource limitations – time and other resources dedicated to examining the problem from stakeholder perspectives and socio-cultural impacts is minimal, and 3) limited self-awareness - default decisions made by coders because of assumptions from their own socio-cultural background or the general culture of the environment around them. Questions around how these ideas can be captured in an accessible process for SED designers represents another front of future work.

Future efforts will also focus on how to connect this work into design processes. In this paper, a post-mortem analysis is used to explore framework design and visualization strategies. In subsequent work, the framework will be adapted to support SED work as it is unfolding. This direction will test how an SED-specific framework can become an effective and accessible design tool.

Given that raw data is an oxymoron, that data is increasingly being integrated in design work, and that SED work is increasingly imbued with sociocultural agency, methods for discussing and assessing critical issues around Soft Materials, from a designer’s perspective, are urgently needed. This is an essential early step in achieving a more complete engagement with the computational components embedded within SED work.

---


This paper was presented at CAADRIA 2018 in Beijing and was first printed in: Learning, Adapting and Prototyping, Proceedings of the 23rd International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2018, eds T. Fukada, W. Huang, P. Janssen, K. Crolla, S. Alhadidi. (The Association for Computer-Aided Architectural Design Research in Asia (CAADRIA), Hong Kong): 2-267-276

Acknowledgements

OnTheLine was developed by DATALab at the School of Architecture at the University of Waterloo in 2014. Project Leads: Mona El Khafif and Maya Przybylski. Student Design Team: Zak Fish, Lea Koch, Daniel Malka, Thomas Noussis and Jake Read.

Maya Przybylski is a Toronto based designer and educator. She is Assistant Professor at the School of Architecture at the University of Waterloo. Maya is a graduate of the Faculty of Architecture, Landscape and Design at the University of Toronto (M.Arch. 2007) where she was awarded the Royal Architectural Institute of Canada Medal for her thesis work on post-oil opportunities for the Caspian Sea. Maya previously earned a degree with a specialization in Software Engineering at the Department of Computer Science at the University of Toronto (B.Sc.Hon. 2003). Through her teaching, research and design work Maya is exploring the complex intersection of architectural design, data and computation. She is developing methods for achieving a more complete engagement, including sociocultural agency as well as technical capacity, with the computational components embedded within data-driven design work. Maya is the director of the DATALab (www.datalab.uwaterloo.ca) research group at the School of Architecture and has collaborated with a range of design practices including: RVTR, Lateral Office, WilliamsonWilliamson, and Bruce Mau Design. She is a founding editor of Bracket (www.brkt.org), a publication series that, since 2010, has documented the intersection of architecture, environment and digital culture and co-author of Pamphlet Architecture 30: Coupling — Strategies for Infrastructural Opportunism (Princeton Architectural Press 2009).