The meaning of 'other' in classifications[1]

Formal methods meet artistic research

Earlier this year, I was invited to enter in a dialogue with Brussels-based artist Rossella Biscotti for the occasion of the exhibition of her installation "Other" from 2015 at the Contour Biennale in Mechelen (Belgium). In this work, she uses the Jacquard weaving technique to visualise data from Belgian census data, and engages in an exploration of data-subjects that are categorised as 'other' within this data-set. The resulting installation consists of 4 large fabrics that display data of various minority-groups and rest-categories of the Brussels population.
My role in this collaboration was to contribute formal or mathematical insights on how rest-categories like *other* or *none of the above* could be understood. In this short piece, I reflect on this collaboration. I first discuss how artistic research like Biscotti's can contribute to the critical evaluation of contemporary data-practices, and then elaborate on how logico-mathematical insights can become part of such inquiries.

Biscotti's 10×10 installation, the precursor of Other, was originally designed and produced to be exhibited at Haus Esters in Krefeld (Germany)—a modernist villa designed by Ludwig Mies van der Rohe for the silk-manufacturer Josef Esters—, and integrates multiple modernist ideals in a single work of art (Holzhey 2014). In this work, Biscotti explores how institutional structures are imposed on individuals by combining features of automated mechanical manufacturing with conceptual and technological aspects of how large data-sets are collected and processed. She focuses in particular on how categories are used to create an overarching structure, and relates this to the punch-cards used to implement such structures within industrial (the Jacquard loom; an early 19th century device that automated the weaving of several complex patterns) and administrative (the Hollerith tabulator) processes that became increasingly automated in the early 20th Century. By showing the resulting work in Haus Esters, it becomes part of a more encompassing modernist narrative exemplified by Mies van der Rohe's architecture.

© Lavinia Wouters

[1] This commentary is an extended version of blog-posts previously published on www.logicandinformation.be and www.oii.ox.ac.uk/blog/tag/logivis/.

For the exhibition of this installation at Contour (the photograph above), Biscotti's team wished to extend their research with a more rigorous expression of the logic behind uses of rest-categories like *other*, and capture this logic in a single formal expression. This led us to an excursion into the meaning of the labels we use to designate such rest-categories, and suggested that we should interpret these labels as semantically empty labels that share certain features with the sentinel values that data-scientists now use to signal that certain data are missing. By asking how such empty labels interact with the generation (and ensuing reification) of categories, for instance when data are aggregated, we came to an interpretation of rest-categories as sets of data-subjects whose members should not, due to the lack of positive evidence of their similarity, be subsumed under a single kind or profile.

The recurrent attention for minority-groups and rest-categories, as well as the value accorded to automated and/or mechanical processes, naturally place Biscotti's work within the scope of current debates on large-scale data-processing and the data-revolution. Mechanical objectivity (Daston and Galison 1992; Christin 2016) and data-shadows are, for instance, current topics of interest within the scholarly community that tries to understand and assess the ethical, legal, and social implications of the data-revolution. In a recent special issue (Leonelli, Rappert, and Davies 2016), several authors interrogate the increasingly influential open data movements by asking how it deals with the many ways in which data can be absent (missing, unreliable, ignored, uncategorised). Such investigations are part of a more encompassing critical enterprise that emphasises that data isn't just what is *given*, but is necessarily also the outcome of a decision to record, store, and eventually redistribute (Kitchin 2014: 2).

The artistic research that led to 10×10 and Other, however, only investigates historical computational technologies like the punch-card, and remains focused on the functioning of categories in census-data, which is itself a very traditional form of large-scale data-collection and organisation that is often associated with the emergence of classificatory and statistical ways of understanding society.[2] It is, therefore, not immediately clear whether and how Biscotti's work, which, unlike the work showed at last year's *Big Bang Data* at Somerset House in London, (Tempini and Leonelli 2016), remains silent on matters like Big Data and machine learning, can contribute (or even intends to contribute) to our understanding of what we now see as the most salient features of the data-revolution.

What I'd like to suggest is that taking early manifestations of automated data-processes as an object of study can help us to open up new ways of questioning data-centric forms of knowledge-production, for instance by making us aware of practices that have become too familiar to deserve a critical assessment. Punch-cards and tabulators are, in that sense, similar to pre-cinematic processes (Braun 1992): they are basic mechanical devices we study to understand the technologies that, respectively, enable contemporary artistic and documentary practices (cinema) or that enable novel epistemic practices. As such, it (re)directs our attention to the technological changes that make epistemic practices possible, or even just conceivable. It becomes a genealogical project, and has the potential to identify the technical and conceptual changes we need to be aware of to understand contemporary practices, by exposing us again to the historical building blocks of our current practices.

[2] "The enthusiasm for numerical data is reflected by the United States census. (…) Printing numbers was a surface effect. Behind it lay new technologies for classifying and enumerating, and new bureaucracies with the authority and continuity to deploy the technology. (…) Categories had to be invented into which people could conveniently fall in order to be counted. (…) Statistical laws that look like brute, irreducible facts were first found in human affairs, but they could be noticed only after social phenomena had been enumerated, tabulated and made public." (Hacking 1990: 2–3).

Biscotti's work helps us, at the same time, avoid certain distractions. It can encourage us to look underneath the reigning rhetoric on Big Data, the mythical abilities that are often attributed to machine learning and artificial intelligence, and perhaps even the most rudimentary principles of inferential statistics. It invites us to take a few steps back—back into what we think of as known territory—, and draws our attention to the practices and assumptions that make data-driven inquiry and decision-making possible: recording, organising and processing through counting, categorisation, and automated calculation.

Because artistic research like Biscotti's is situated at the periphery of current scholarly debates, it isn't bound by a given research-agenda and can reinvestigate familiar and often widely trusted practices, and ask elementary questions anew; from a contemporary (artistic) perspective.[3] By integrating unexpected perspectives and easily overlooked cultural and historical references, an artistic intervention (in this case, an installation, but also the research that precede its conception and elaboration) can be used to raise questions that may have lost their immediate relevance because they no longer drive our scientific or scholarly curiosity, but also questions that are not aligned with the dominant themes of ongoing debates concerning privacy, fairness, transparency, or responsibility.

What then can a logico-mathematical approach contribute to artistic research concerned with the classification practices on which census-data are built? Two things at least. It can help make the idea of a "logic of classification" more explicit, and develop its implications in purely abstract terms (for instance without associating rest-categories with forms of exclusion). As such, it can reorient our critical attention from how classification-structures affect specific data-subjects in concrete settings to how classification-rules create abstract entities like the profiles or categories that become the primary entities we reason about or use to make decisions. Second, it can be used to explore alternative approaches; in this specific case, different ways of conceptualising how rest-categories should be used in the construction of categories of (in certain respects) similar data-subjects.

In its most basic incarnation, categorisation can be understood in terms of two epistemic actions. Categorisation first requires an abstraction-step (Floridi 2008; Victor 2011); a decision to take into account some properties of individuals, and to ignore some others. As such, we create groups of data-subjects that are identical (or sufficiently similar) along a given selection of data-dimensions. At this point, the individuals that end up in the same category become somehow interchangeable: they may contribute to some aggregated properties associated with a category of individuals, but their specific contribution remains hidden (Stigler 2016: Chapter 1). This process may be followed by a *reification* step. This move replaces groups of individuals by the categories themselves, which then become the first-class citizens of the ontology we use to manipulate data from a given population.

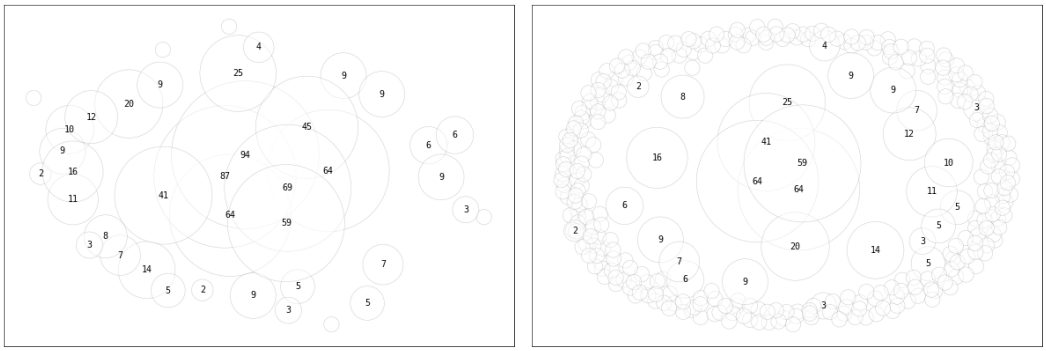
In relation to the focus on "other", I specifically contrasted two different ways in which the membership of a rest-category could be conceptualised. The basic principle that underlies both is that data-subjects belong to the same category (or fall under the same profile) if and only if for all the relevant data-dimensions we have attributed them the same values (or values within the same range). In this way, we can construct categories of, say, all the children of

[3] In a lecture given at the 2012 summer school on "Images and Visualisation: Imaging Technology, Truth and Trust," in Norrköping, Sweden Anne Beaulieu makes a similar point with respect to how artists in residence can force other kinds of discussions, and challenge existing conventions within information visualisation.

ages between 6 and 10 that have at least one sibling. Similarly, we seem to be able to construct the category of all the data-subjects categorised as “other” in the data-dimension “household position,” and this even if the actual household-roles of the presumed members of this category do not have anything in common apart from the fact that they do not conform to any of the roles privileged by the designers of the census, and that their place or role within a household probably isn’t very common (as in the case of “other nationalities”). Treating such rest-categories as *bona fide* categories makes sense if we think of labels like “other” or “none of the above” as semantically significant labels; labels that provide sufficient ground for identification because they indicate that we have sufficient evidence to identify the data-subjects that were so-labelled, and perhaps even extend our ontology with a corresponding category.

If, however, we think of such labels as a mere indication of the absence of any information, this strategy quickly becomes questionable. In the context of the mentioned household positions, being categorised as “other” results from negative answers to 4 consecutive yes/no-questions (Holzhey 2014: 49), but does not need to carry any positive information. At least for some rest-categories it thus makes more sense to treat the labels we use to denote these categories along the same lines of the sentinel-values that are customarily used to signal missing data, like 9999 or the `NaN` (not a number) numeric data-type described by the IEEE 754 floating-point standard. Let us stipulate that two data-subjects fall under the same profile or belong to the same category if and only if, *first*, there is no information that indicates that they are different in a relevant respect (a potentially vacuous sense of being similar), and, *in addition*, there is also positive evidence that they are similar in the relevant respects. By the second requirement, the label “other” `then` no longer leads to the creation of a category of *others*. Because explicit sentinel-values like `NaN` have the property of not being equal to themselves (the expression `NaN==NaN` will typically evaluate to `False`), this requirement for positive information can be simulated by using such values to denote rest-categories.

Using a randomly generated data-set similar to the data used by Biscotti, the difference between the two types of approaches can easily be visualised. In the figures below the sizes of categories are displayed as bubbles; the figure on the left uses the number 10 to denote “other” (and `10==10` evaluates to `True`), whereas the figure on the right uses `NaN`.



Here, we immediately see that the presence of data-subjects labelled as “other” leads to the creation of a large periphery of different (because unknown) data-subjects whenever the label used to denote rest-categories indicates the absence of information. As such, this leads to a

minimal sense in which we can understand how the meaning we assign to the labels we use to denote categories interacts with the process of creating categories or profiles and the subsequent use of these categories as an ontology used to describe a given subject-matter.

References

Braun, Marta. 1992. *Picturing Time: The Work of Etienne-Jules Marey (1830-1904)*. University of Chicago Press.

Christin, Angèle. 2016. "From Daguerreotypes to Algorithms: Machines, Expertise, and Three Forms of Objectivity." *SIGCAS Comput. Soc.* 46 (1). New York, NY, USA: ACM: 27–32. doi:10.1145/2908216.2908220.

Daston, Lorraine, and Peter Galison. 1992. "The Image of Objectivity." *Representations* 40: 81–128.

Floridi, Luciano. 2008. "The Method of Levels of Abstraction." *Minds and Machines* 18 (3): 303–29.

Hacking, Ian. 1990. *The Taming of Chance*. Cambridge: Cambridge University Press.

Holzhey, Magdalena R, ed. 2014. "Rossella Biscotti — 10 X 10." In . Kunstmuseum Krefeld / Museum Haus Esters — Revolver Publishing.

Kitchin, Rob. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Sage.

Leonelli, Sabina, Brian Rappert, and Gail Davies. 2016. "Data Shadows: Knowledge, Openness, and Absence." *Science, Technology, & Human Values*. SAGE Publications Sage CA: Los Angeles, CA, 162243916687039.

Stigler, Stephen M. 2016. *The Seven Pillars of Statistical Wisdom*.

Tempini, Niccolo, and Sabina Leonelli. 2016. "Exhibition: Live by Data, Die by Data." *Nat Phys* 12 (2). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 109–10. http://dx.doi.org/10.1038/nphys3652.

Victor, Bret. 2011. "Up and Down the Ladder of Abstraction." http://worrydream.com/LadderOfAbstraction/.

Patrick Allo
Oxford Internet Institute
University of Oxford

http://www.oii.ox.ac.uk
http://digitalethicslab.oii.ox.ac.uk
http://www.logicandinformation.be