2011

First-Class Objects

James Grimmelmann, *New York Law School*

FIRST-CLASS OBJECTS

JAMES GRIMMELMANN*

"You are not special. You are not a beautiful or unique snowflake."[1]

I. UNIQUE IDENTIFIERS

How should we think about privacy in a digital age? One approach is to focus on how people use computers: how what we choose to share about ourselves changes when we go online.[2] But we could also focus on how computers use people: how flows of personal information are transformed by technology. Just as email is different from mail, a spycam is different from a spy.

This brief essay will examine a seemingly technical question: *how are people represented within computer systems*? The essay will argue that that there are two possible ways to do it, and that the choice between them has important technical, social, and humanistic consequences. It won't

[*] Associate Professor of Law, New York Law School. I presented earlier versions of this paper at the Privacy and Innovation Symposium at Yale Law School and the Privacy and the Press conference at the University of Colorado Law School. My thanks for their comments to the attendees there, to the participants in the MSRNE Social Media Collective mailing list, and to Aislinn Black, Frank Pasquale, and Tal Zarsky. This essay is available for reuse under the Creative Commons Attribution 3.0 United States license, http://creativecommons.org/licenses/by/3.0/us.

[1] Fight Club (Fox 1999).

[2] *See, e.g.*, James Grimmelmann, *Saving Facebook*, 94 IOWA L. REV. 1137, 1147-50 (2009).

say much new about those consequences—instead, it will show how closely linked they are.

A. James Grimmelmann and @grimmelm

The difference is illustrated by a tweet. On October 27, Ryan Calo sent the following text to Twitter:

> Privacy and innovation thought pieces by Helen Nissenbaum, Frank Pasquale, @grimmelm, and others up on Yale ISP. http://bit.ly/aUtk0v[3]

Let's examine two parts of this tweet: "Frank Pasquale" and "@grimmelm".[4] Syntactically, they're both strings of characters from the Latin alphabet, enriched with some standard punctuation symbols. They contain 14 and 9 characters, respectively. In the standard UTF-8 encoding used by Twitter,[5] they would take up 14 and 9 bytes, that is, 112 and 72 individual ones and zeros.[6]

Semantically, "Frank Pasquale" and "@grimmelm" are both names; their preferred interpretation is that they refer to people. "Frank Pasquale" is what Calo typed so that readers of his tweet would know he was talking about Frank Pasquale, the Schering-Plough Professor in Health Care Regulation and Enforcement at Seton Hall Law School. "@grimmelm" is what Calo typed so that readers would know he was talking about me.

This second meaning requires some explanation. "grimmelm" is my Twitter username, so "@grimmelm" is a way of referring to me. Since Twitter limits all posts to 140 characters, space is at a premium, and concision is essential. In 2007, Twitter user Chris Messina started using the pound symbol "#" to flag the topics of his tweets, such as "#barcamp" for a message of interest to attendees of the Bar Camp event.[7] These "hashtags" caught on, and millions of Twitter users began deploying them to annotate a wide range of tweets. The Twitter community embraced other compressed forms, such as the dollar sign "$" followed by

[3] Ryan Calo, TWITTER (Oct. 27, 2010, 12:24 AM), http://twitter.com/#!/rcalo/status/28913525284.

[4] Here, and throughout this essay, I have moved punctuation marks outside of quotations for purposes of precision.

[5] *See Counting Characters*, TWITTER DEVELOPERS, http://dev.twitter.com/pages/counting_characters (last visited Feb. 24, 2011).

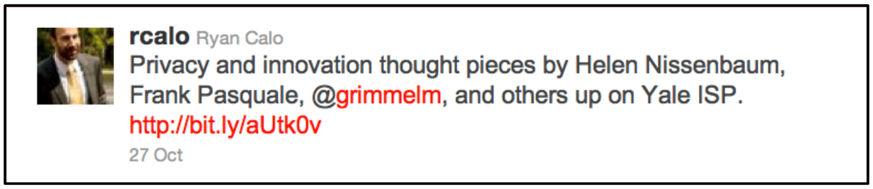
[6] *See* UNICODE CONSORTIUM, THE UNICODE STANDARD 3-5 (5th ed. 2006); JUKKA K. KORPELA, UNICODE EXPLAINED 298–300 (2006).

[7] *See* Chris Messina, *Groups for Twitter; or A Proposal for Twitter Tag Channels*, FACTORYCITY (Aug. 25, 2007, 10:00 PM), http://factoryjoe.com/blog/2007/08/25/groups-for-twitter-or-a-proposal-for- twitter-tag-channels.

a stock ticker symbol to refer to a company (e.g. “$”[8]), and the at-sign (“@”) followed by a username to refer to a particular user (e.g. “@grimmelm”).[9]

Within the community of Twitter users, “James Grimmelmann” and “@grimmelm” would both be recognized as valid names for me. A Twitter user who knew me would understand that they referred to me; a Twitter user who didn’t know me would still surmise that they referred to someone with that name or username. These usages are both conventional. True, the tradition of assigning and capitalizing names is older, more widely known, and more universally followed. More people will recognize “James Grimmelmann” than “@grimmelm”. But at root, they are both conventions within an interpretive community of humans.

Twitter, however, treats “Frank Pasquale” and “@grimmelm” differently. Here is what Calo’s tweet looked like on Twitter’s website:



rcalo Ryan Calo
Privacy and innovation thought pieces by Helen Nissenbaum, Frank Pasquale, @grimmelm, and others up on Yale ISP. http://bit.ly/aUtk0v
27 Oct

In the posted version, “Frank Pasquale” appears normally, in black type. “grimmelm”, however, appears in red. That’s because it’s a hyperlink; it links to my profile page on Twitter. Once the @-syntax caught on among users, Twitter adapted to it. The company reprogrammed its software to turn each such string—an “@” followed by a username—into a hyperlink to that user’s profile page on Twitter.com.[10]

This isn’t just a difference between one kind of name and another—“James Grimmelmann” versus “@grimmelm”. It also means that Twitter can make distinctions among users who are named in tweets. Old-school plain-text names, such as “James Grimmelmann” and “Ryan Calo”, are blobs of unstructured data, no different from “d#fh@@3.pQMNa0”. But tweets with “@rcalo” and “@grimmelm” and “@amturing” now have structure; Twitter’s software can and does do different things depending on who is named by the tweet. For example, Twitter now builds for each

[8] *See, e.g.*, Mad Money on CNBC, TWITTER (Feb. 3 2010, 6:11 PM), http://twitter.com/#!/MadMoneyOnCNBC/status/33346776282963969.

[9] *See, e.g.*, David Christiansen, *How to Use @Reply in Twitter Messages*, INFORMATION TECHNOLOGY DARK SIDE (Mar. 11, 2009), http://www.techdarkside.com/how-to-use-reply-in-twitter-messages.

[10] @Ev [Evan Williams], *How @Replies Work on Twitter (and How They Might)*, TWITTER BLOG (May 12, 2008, 10:51 AM), http://blog.twitter.com/2008/05/how-replies-work-on-twitter-and-how.html.

user a list of the tweets that mention her, whoever they were posted by.[11]

B. Unique Identifiers

Twitter's user database is central to its ability to make "@grimmelm" meaningful. If I write "@asdfasdfhjsa" in a tweet and click on the link that results, Twitter will display an error message that says, "This user does not exist." Twitter has a list which records the facts that "grimmelm" and "rcalo" are usernames but that "d#fh@@3.pQMNa0" and "asdfasdfhjsa" are not. In particular, Twitter's database assigns a *unique identifier* to each user. Since unique identifiers are ubiquitous in computer science, it will be helpful to discuss the technical considerations behind them.

Unique identifiers come from the world of databases, in which one seeks to store information about the world in a structured manner.[12] One way of thinking about the problem is that one wants to keep track of things and how they relate to each other. The widely-used "entity-relationship model" formalizes this idea by describing the world as a collection of *entities*.[13] "An entity is a 'thing' which can be distinctly identified. A specific person, company, or event is an example of an entity."[14] One describes the world by specifying the *attributes* that entities have (e.g. "John Doe's height is 5'9") and the *relationships* in which they participate (e.g. "John Doe and Jane Roe are married"). Unique identifiers pervade the model. Not only is an entity defined in terms of its ability to be uniquely identified, but to say that an entity has an attribute or participates in a relationship, one needs to be able to identify the entity in question.

Similar questions arise when one confronts the problem of database design: how best to store a representation of the world in a computer database. The dominant modern database paradigm is the "relational model," in which a database consists of a collection of *tables*.[15] The "table" metaphor is based on the two-dimensional display of tabular data on paper in rows and columns. Each row (or *entry*) consists of a series of

[11] *See What Are @Replies and Mentions?*, TWITTER HELP CENTER, http://support.twitter.com/groups/31-twitter-basics/topics/109-tweets-messages/articles/14023-what-are-replies-and-mentions (last visited Feb. 24, 2011).

[12] *See generally* RAGHU RAMAKRISHNAN & JOHANNES GEHRKE, DATABASE MANAGEMENT SYSTEMS (2d ed. 1999); C.J. DATE, AN INTRODUCTION TO DATABASE SYSTEMS (7th ed. 1999); ABRAHAM SILBERSCHATZ ET AL., DATABASE SYSTEM CONCEPTS (6th ed. 2010).

[13] *See* Peter Pin-Shan Chen, *The Entity-Relationship Model—Toward a Unified View of Data*, 1 ACM TRANSACTIONS ON DATABASE SYS. 9 (1976).

[14] *Id.* at 10.

[15] *See* E.F. Codd, *A Relational Model of Data for Large Shared Data Banks*, 13 COMM. ACM 377 (1970).

values, one each from the categories named by the column headings (or *fields*). For example, a course's table in a registrar's database might store the value "4" where the row for Wills and Trusts intersects the column for number of credits.[16] More concisely and precisely, we would say that the entry for Wills and Trusts has the value "4" in the number-of-credits field.

Here again, unique identifiers are pervasive. They are often necessary if we are to meaningfully combine, or *join,* the information from multiple tables.[17] Unless we have a way to know that the Wills and Trusts in the courses table is the same as the Wills and Trusts in the table of student schedules, there is no way to generate student transcripts with the correct number of credits. Giving Wills and Trusts a unique identifier—a common value that appears in both tables—provides an answer. A large literature on database design deals with the problem of finding or creating identifiers, or *keys*, that suffice to tell different rows apart, and with ensuring that their usage in different tables is consistent enough to permit meaningful joins.[18]

The vital role of unique identifiers for users in Twitter should now be apparent.[19] When a new tweet refers to me or to Ryan Calo using @-syntax, Twitter determines which user it should be associated with by consulting the database. Whenever any new information that should be connected up with a particular user comes in—a password change, a new tweet, a new follower, etc.—that information is added to another table in an entry that also includes that user's unique identifier. Everywhere inside Twitter's systems that a unique identifier goes, it is intended to refer to a specific user, and does.

Users are entries in Twitter's databases; they have unique identifiers. By contrast, for example, musical notes are *not* entries in Twitter's databases. Neither are cities, emotions, galaxies, cars, or judicial opinions. One can talk about these things on Twitter, and much much more, but not in a way that Twitter's servers will understand in the slightest. In contrast, one can talk about Twitter users in a way that Twitter will get; it will know *who* you're talking about, and be able to

[16] The "intersection" is metaphorical, of course.

[17] *See, e.g.*, RAMAKRISHNAN & GEHRKE, *supra* note 12, at 97–98.

[18] *See, e.g.*, DATE, *supra* note 12, at 258–64.

[19] Twitter does not use these character strings as the actual unique identifiers. Instead, because numbers are easier for computers to work with than strings, Twitter assigns each user a unique ID number. *See, e.g.*, *Twitter REST API Method: users show*, TWITTER API DOCUMENTATION, http://apiwiki.twitter.com/w/page/22554755/Twitter-REST-API-Method:-usersshow (last visited Feb. 24, 2011). Whenever it sees a @username in a tweet, Twitter translates it into the appropriate ID number and uses the number internally from then on. *Cf. Find your Twitter ID*, IDFROMUSER.COM, http://www.idfromuser.com (last visited Feb. 24, 2011) (allowing one to look up the corresponding numerical user ID by typing in a Twitter user's screen name).

react accordingly. Unique identifiers are the essential catalyst in transforming messes of unstructured information into useful, structured data about people.

C. Other Examples

This phenomenon is hardly confined to Twitter. Many other computer systems use unique identifiers for people. Consider a few more examples:

Facebook was designed from the ground up to give people unique identifiers. It assembles real names and other personal information into highly-structured profiles linked to a unique user identifier.[20] These profiles can be sorted, searched, and automatically manipulated. Try clicking on a favorite movie in a friend's profile, for example, and part of the resulting page will contain a list of your friends who also picked that movie as a favorite—a computation that joins multiple database tables (your friends, favorite movies) by matching unique identifiers. Facebook is thus profoundly oriented towards associating information with people: it collates, categorizes, analyzes, exposes, and projects them.

Another classic example of databases in which entries represent people is the credit reporting agency.[21] In order to report a credit history or credit score for a person, the agency must maintain a file for that person. This file takes the form of a unique identifier that is then cross-linked in a database with every transactional datum available on the person to whom that identifier corresponds: mortgage payments, credit card limits, past addresses, and much much more. Social Security numbers have traditionally been the unique identifier of choice, but due to fraud and mistakes, they're not always entirely reliable.

By way of contrast, consider the Wayback Machine's near-comprehensive archive of the Web.[22] It crawls the Web repeatedly, taking snapshots of every webpage it finds. Users can then retrieve a historical archive of any given webpage, seeing what it looked like on various dates stretching out across years. Many of these pages refer to people. When they do, however, the Internet Archive has no idea that they do. Names are just blobs of text, indistinguishable from any of the other blobs of text in the archived webpages. People are *not* entries in the Internet Archive's databases.[23]

[20] *See, e.g.*, Grimmelmann, *supra* note 2.

[21] *See, e.g.*, Daniel J. Solove, *Privacy and Power: Computer Databases and Metaphors for Information Privacy*, 53 STAN. L. REV. 1393, 1408 (2001).

[22] INTERNET ARCHIVE WAYBACK MACHINE, http://waybackmachine.org (last visited Feb. 24, 2011).

[23] If I retrieved pages from the Wayback Machine and then scanned them for text that looked likely to be a name, I might create a system that had identifiers for people, but the

II. CONSEQUENCES

Let us explore some of the consequences of giving people unique identifiers in order to create database entries on them. This simple technical move has surprisingly wide-ranging effects. It connects to so many observations in privacy and technology scholarship that it suggests there is something fundamental about the shift. Unique identifiers are the key, so to speak, to the process by which computer systems become *about* people.

A. Standardization

Unique identifiers and structured data are inherently *standardized*. By imposing structure, one can produce a well-defined representation that is free from much of the ambiguity of unstructured data. As we shall see, this standardization is central to the tremendous power of unique identifiers. But since the world is itself unstructured and ambiguous, the process of standardizing identifiers introduces its own errors. I will break standardization down into four components: *uniqueness* of identifiers, *normalization* of them to give people canonical names, the inevitable *errors* that result, and the discontinuous way in which data attached to unique identifiers *decays*.

Ordinary names aren't unique: think of "John Smith".[24] Compare that with Twitter usernames: there is only one "@grimmelm". The "unique" in "unique identifier" requires that different people have different identifiers. The flipside of uniqueness is normalization. Sometimes people call me "Jim", which isn't quite right—but isn't quite wrong, either. These slippages are unproblematic in everyday life, but the kind of contextual insights people bring to the table are hard for computers to replicate. Unique identifiers deal with the problem by making identifiers canonical. Instead of dithering over whether I prefer to be called "James" or "Jim", just use "@grimmelm". It does the right thing.

Getting to @grimmelm, however, isn't as easy as it looks. The first problem is inherent in the need for uniqueness: the real world is filled with people who use identical or confusingly similar names. Precisely because there can be only one "@grimmelm", only one of us can have it, and that means conflict. The endemic and enduring fights over domain names[25] are echoed in the land-rush every time a new social media service

process would be imperfect, approximate, and error-prone.

[24] *See* James Gleick, *Get Out of My Namespace*, N.Y. TIMES MAG., Mar. 21, 2004, at 44 ("You don't own your name. Just ask any John Smith.").

[25] *See generally* JACQUELINE LIPTON, INTERNET DOMAIN NAMES, TRADEMARKS, AND FREE SPEECH (2010).

hands out identities on a first-come, first-served basis.[26] Even using artificial identifiers can be a technical challenge: they need non-trivial infrastructure to create, distribute, and manage.[27] Name assignment is inherently political.[28]

The second problem is that while a set of unique identifiers may be clean and well-structured, the world is anything but. The process of mapping the world onto those identifiers can never be specified completely and correctly. Someone has to enter the data; that someone will make typos and bad guesses. Whenever data from two different databases or sources is to be combined (which is quite often, as unique identifiers make this aggregation attractive), mismatches between their identifiers introduce fresh errors. Identity theft, wrong addresses, conflation with other people with the same name—all of these crossed wires can be triggered when a credit file is populated with outside information which is mistakenly assigned to your identifier in the database. In database terminology, these mistakes are the results of an improperly specified join operation—one that combines two tables using a poorly-chosen key.

Another source of error is the passage of time, and here, structured data is a mixed blessing. One the one hand, standardization plays a centripetal role by facilitating error correction. Misspellings and other minor mistakes are easier to spot and repair before they cascade and feed each other. On the other hand, digitization and centralization increase the risk of truly catastrophic failure. For example, when the servers supporting Microsoft's Sidekick mobile phone customers failed, thousands of users suddenly lost access to their contact books.[29] The price we pay for resilience against daily small errors is a greater risk of a single big failure.

Standardization helps here: many or most random errors become easily-spotted syntactic mistakes (think of how much faster it is to spell-check a word processing document than the same manuscript in printed form). Normalization plays a centripetal role, fixing up misspellings and eliminating other minor mistakes before they multiply and feed each other. On the other hand, this centralization increases the risk of truly catastrophic failure.

[26] *See, e.g.*, Verne Kopytoff, *Facebook Land Rush to Start in Three Days*, TECH CHRONS. (June 9, 2009, 3:45 PM), http://www.sfgate.com/cgi-bin/blogs/techchron/detail?entry_id=41455.

[27] *See, e.g.*, INT'L TELECOMMS. UNION, STANDARD X.667 (2004) (34-page international standard on generating and distributing unique identifiers).

[28] *See generally* MILTON MUELLER, RULING THE ROOT: INTERNET GOVERNANCE AND THE TAMING OF CYBERSPACE 87-88 (2004).

[29] *See* Rob Pegoraro, *Sidekick Users See Their Data Vanish Into a Cloud*, WASH. POST, Oct. 13, 2009, at A14. Sidekick users' information, such as address books and to-do-lists, was primarily stored on company servers. Maintenance of the servers went wrong, and backups proved unusuable, locking users out from their data.

B. Third Parties

Unique identifiers don't just happen on their own. Someone has to build the database, create identifiers, and ensure that they really are unique. The use of unique identifiers, in other words, is inherently tied to particular *third parties*. "@grimmelm" has its special meaning because of Twitter's efforts. Similarly, you need to consult a credit agency's files to run a credit report on someone, and social security numbers depend on the Social Security Administration's coordinating role. Without these third parties, unique identifiers lose their special meanings. If Twitter vanished tomorrow, "@grimmelm" would become an ordinary name again, like "James Grimmelmann". People could still use it to refer to me, but this would be a matter of convention and tacit human knowledge, not an automated, fixed reference.

We will have much more to say about third parties, but for the moment, I would like to emphasize two ways in which their special role manifests itself: the *dependence* users have on the third party's continued support, and the *lock-in* the third party enjoys against user attempts to switch to another third party. Dependence first: The more valuable and important an identifier, the more one has to lose if it goes away. Because a unique identifier is controlled by a specific entity, rather than being dispersed throughout a community, as a traditional name would be, one becomes dependent on the entity. The third party who holds a unique identifier holds the name itself hostage, and possibly the person. As anyone who's been locked out of their email account can attest, losing an important unique identifier can be devastating. If Facebook collapses, all the information locked in its proprietary formats and adapted to its social network will be simply gone.

These third parties also enjoy a kind of lock-in effect, precisely because *other people* use them to interact with and learn about you. No one wants to be the only person on a social network; no one would query a credit agency with a single file. But if everyone in your industry is on LinkedIn, you may need to be too, and if every landlord uses the same background-check service, you had better worry about what your file says about you.

Compounding the problem, it's much harder to move structured data around than unstructured data. To leave Facebook for a competing social network, for example, I will need to export the data in a structured format (which Facebook does not currently allow or enable), and find a competitor using a compatible format for its own data.[30] Then there is

[30] *See* Robert Scoble, *Facebook Has a Point Where It Comes to Your Privacy*, SCOBLEIZER (May 15, 2008), http://scobleizer.com/2008/05/15/facebook-has-a-point-where-it-comes-to-your-privacy.

the problem of interoperability: for example, Facebook now provides login services for other websites and services, including Skype. One could see this either as making identity more portable by allowing a user to sign in only to Facebook, or as making identity less portable by forcing everything to flow through Facebook.

C. *Knowledge Creation*

Using unique identifiers for people enables a wide variety of practices that involve the *creation of knowledge* about them. I will bring out four, which build on each other: the *aggregation* of information about a person from multiple sources, *automated reasoning* about a person from multiple pieces of information, the *enumeration* of all of the references to a given person in a database, and *statistical analysis* about populations by summarizing information about multiple people.

Unique identifiers are remarkably convenient focal points for data aggregation. Within a database, this is often the point of having unique identifiers at all: to allow them to serve as keys for joining data from two different tables. That works with unique identifiers; it doesn't work without them. The registrar can put information about my courses from the courses table together with information about me from the faculty table to produce a personalized schedule that indicates when I am expected to be in class. There is more information in this combined view than there was in either table alone. This same phenomenon can happen on a larger scale when multiple databases are brought together—or when new information is added to an existing database. Having entries for people in a database is an essential step in bringing together information about them from many different sources.

Once multiple pieces of information are associated with a person in a database, it becomes possible to ask a computer program to engage in automated reasoning about them. A credit score is one kind of automated reasoning: one that results from algorithmically combining large quantities of financial data according to a set formula. Similarly, Foursquare can conclude that multiple people are in the same physical space based on their separate check-ins, and Amazon can recommend new books based on previous ones you've purchased, viewed, and reviewed. This is the Semantic Web dream, of course: everything encoded in a way that supports the creation of complex relationships of out simpler pieces—that is, drawing conclusions on the basis of aggregated data.[31]

One particularly simple, but important, form of automated reasoning is *enumeration*: listing all of the references to a given person.

[31] *See* Tim Berners-Lee et al., *The Semantic Web*, SCI. AM., May 2001, at 34, 36-38.

That is, you can look through Twitter for all the tweets that mention a user or through Facebook for all the photos in which someone is tagged, and have high confidence that you have seen all such items that are possible for you to see. This property depends on normalization and the use of third parties. The third party is a single source maintaining a complete list of data about a person, and normalization means there is a standard way of ensuring that all references to that person are associated with their digital identifier. This property doesn't hold in general; I am quite certain that I don't know all of the places I'm referenced on the Web. For a lawyer doing due diligence, a private investigator building a file, or a nervous college student untagging photos of herself at a keg party, enumeration is a godsend.

A different way to extend automated reasoning is to draw conclusions not about individuals but about populations. This is the goal of statistical reasoning. Here, the use of unique identifiers reaches back far beyond the dawn of digital computing, into the parallel growth of bureaucracy and demography. The data miner deciding which customers are most likely to respond to a promotional flyer for a new toothpaste, the transportation planner estimating the number of subway cars needed over the next five fiscal years, and the pollster gauging support for a candidate are all dealing with abstracted statements about people. The unique identifiers may have receded into the background here, but note that these exercises are futile unless they start by identifying and differentiating the characteristics of individuals. Gauging the likely outcome of an election by surveying the same person five hundred times is ridiculous; surveying five hundred different people is not.

D. Representation

We have noted that unique identifiers are essential for representing people in databases. But there is another kind of *representation* that they enable: to other people. Unique identifiers are pervasively linked to social uses of digital technology, because they play all sorts of roles in shaping the presentations of people that other people see. I would like to call out four in particular: voluntary *self-presentation* by shaping how one's digital persona is built up, increased and involuntary *visibility* of one's actions and attributes, the possibility of *misrepresentation* of a person by a distorted digital persona, and proactive *monitoring* of one's digital presence.

How does an online persona differ from the numerous offline personas people have always created for particular social roles? A unique identifier provides a centralizing, coordinating location for aggregating various personal qualities into the digital self one wishes to show to the world: an email address or a social network profile. Beyond that, though, people seem almost to gravitate to using structured data for their self-identification. From the Geek Code to the well-defined slots in a Facebook profile to the millions of online quizzes people fill out to tell others about themselves,[32] there seems to be a natural enthusiasm for crafting digital avatars using well-defined categories.[33] It may have to do with the creativity-promoting qualities of constraint, but also with the social usefulness of structured signals. A unique identifier provides the fixed point to which these additional attributes can be attached in a structured way.

On the other hand, if you're in a database, it's harder to hide. You're more visible, because data sticks to unique identifiers like cat hair to sweaters. We all know about the gigantic databases that commercial profilers have on all of us. These identifiers also help stalkers and other private individuals do the same. If I'm trying to look you up, I can get much further once I figure out what your Twitter handle is. You may not have put your real name on the account, but if I can infer that it's you, the centralized, normalized role that it plays helps me build an extensive file on you quickly. It is no accident that thinkers have cast about for metaphors to express the uniquely personal, uniquely threatening characteristics of these new databases: Daniel Solove's "digital dossiers,"[34] John Battelle's "database of intentions,"[35] Paul Ohm's "databases of ruin."[36]

Moreover, visible data need not be correct data; we have already noted the pervasiveness of errors in databases about people. Not only can anyone who supplies data about a person get it wrong, but the third parties who control the unique identifiers have a special kind of power over how a person is represented. Just as a credit rating agency can destroy my ability to get a mortgage, Facebook could metaphorically scribble a mustache on my profile or Twitter could redirect every mention of "@grimmelm" to my mortal enemy.

With enumeration, however, comes the possibility that one could protect one's privacy through proactive monitoring. If you want to keep something secret, but there are many places where people could be talking about the secret, then you have a Pokemon problem: gotta catch 'em all. It is much easier for you to make that search when you can

[32] *See, e.g.*, Which Hegel's Phenomenology of Spirit Character Are You? SELECTSMART.COM, (October 2000), http://www.selectsmart.com/FREE/select.php?client=hupitesti.

[33] *See* Grimmelmann, *supra* note 2, at 1176.

[34] DANIEL SOLOVE, THE DIGITAL PERSON: TECHNOLOGY AND PRIVACY IN THE INFORMATION AGE 2 (2006).

[35] JOHN BATTELLE, THE SEARCH: HOW GOOGLE AND ITS RIVALS REWROTE THE RULES OF BUISNESS AND TRANSFORMED OUR CULTURE 1-2 (2005).

[36] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1705 (2010).

enumerate *every* reference to yourself. Facebook, for example, sends me a message every time someone tags a photo of me, and lets me refuse entirely to be tagged in Places.[37] This works only imperfectly in a non-centralized, non-normalized space like the Web. It leaves me dependent, however, on the good will of the third party to let me step through all of the relevant references. If it hides the references from me, I can do nothing.

E. Control

Finally and most famously, unique identifiers profoundly shape the dynamics of power and *control* around personal information. Some of these moves empower individuals; others leave them comparatively helpless. I will bring out four themes from these extensive debates, all of which can be linked to unique identifiers: *empowerment* of individuals by helping them accumulate self-knowledge in a structured way, panoptic *control* of individuals by outside entities who use the identifier as a focal point, *manipulation* of individuals by those who use detailed personal profiles to shape what they see and think, and the pure existential *objectification* of individuals by others who "reduce" them to an entry in a database.

Start with empowerment: Having structured data about oneself in digital form can be useful. The electronic health record is probably the best example. It's enormously helpful for me to have a single digital file that I can share with a new doctor, rather than there being scattered information about me in different locations, digital and paper, which I have to search out, pore through, and compile. This is the positive face of aggregation. Lifelogging is a kind of self-help version of aggregation with precisely this goal: collecting and collating large quantities of data about oneself to grow in self-knowledge.[38]

On the other hand, knowledge is power. Governments have known this since long before the digital age. The Domesday Book and the secret police file catalogue information on people and their activities. The census, from the age of the punch card on, added database structure. These are the tools of rational administration, the essential inputs to bureaucracy and the extension of governmental power. On the one hand, this facilitates technocratic expert administration; on the other hand, punch-card technology helped organize the deportation and execution of Jews during the Holocaust.[39] Sorting people based on their

[37] *See Privacy, Editing, Tagging, and Abuse*, FACEBOOK HELP CENTER, http://www.facebook.com/help/?page=831 (last visited Mar. 11, 2011).

[38] *See* GORDON BELL & JIM GEMMELL, TOTAL RECALL: HOW THE E-MEMORY REVOLUTION WILL CHANGE EVERYTHING 127 (2009).

[39] *See* EDWIN BLACK, IBM AND THE HOLOCAUST: THE STRATEGIC ALLIANCE

characteristics is a form of comprehensive control over them.[40]

The fear of control based on personal characteristics is also central to debates over personalized, targeted advertising.[41] What some authors see as empowerment, others see as manipulation. The advertising firm that builds a profile of your browsing habits (even, perhaps, if it can't identify you by name) nonetheless uses that personal profile to mark you and market to you. It uses that knowledge—which is made specific and actionable by the database entry—to exert power over you, possibly to your disadvantage.

Finally, some go a step further and argue that being represented in a database can be intrinsically objectifying. It flattens out one's identity to the standardized forms supported by the system. When protesters marched against computerization in the 1960s, with shouts that people were not to be "folded, spindled, or mutilated," this was the idea at work.[42] It is possible to argue that being represented in a database is intrinsically demeaning to one's human dignity. It strips out the respect for your personhood that demands you be recognized as a full, worthy, complex *person*, not just a reductive set of binary digits.

CONCLUSION

This has been an essay about representing people in databases. I have argued that the transition from unstructured data to structured data is of critical importance for thinking about privacy and social interactions. There are echoes of at least three previous shifts in this transition: the introduction of print, the growth of bureaucracy, and the rise of digital media. All three of them have reworked the relationships of individuals to each other, and to the larger institutions that make up their worlds: communities, companies, and countries. The use of unique identifiers as the keys to structured databases about people will have its own dramatic consequences.

Another computer science term, this one from the field of programming languages, is suggestive of the values at stake. One sometimes says that a system which directly represents certain things treats them as "first-class objects."[43] One computing website explains

BETWEEN NAZI GERMANY AND AMERICA'S MOST POWERFUL CORPORATION 44 (2001).

[40] *See* OSCAR H. GANDY JR., THE PANOPTIC SORT: A POLITICAL ECONOMY OF PERSONAL INFORMATION 134 (1993).

[41] *See, e.g.*, Tal Z. Zarsky, "*Mine Your Own Business!": Making the Case for the Implications of the Data Mining of Personal Information in the Forum of Public Opinion,* 5 YALE J.L. & TECH. 1, 50-53 (2002).

[42] *See* Steven Lubar, *"Do Not Fold, Spindle or Mutilate": A Cultural History of the Punch Card*, 15 J. AMER. CULTURE 43, 46-48 (2004).

[43] *See* MICHAEL L. SCOTT, PROGRAMMING LANGUAGE PRAGMATICS 141 (2d ed. 2006); HAROLD ABELSON ET AL., STRUCTURE AND INTERPRETATION OF COMPUTER

that an element in a programming language is first-class "when there are no restrictions on how it can be created and used."[44]

For example, in some programming languages, like C, functions are not first-class. Any subcomputation that the program will carry out must be specified in advance by the programmer, and there are significant limits on how functions can be stored, modified, and passed around. In other programming languages, like Scheme, functions are first-class: the computer treats them just like it would any other kind of data, like a number or a binary true/false. This leads to great flexibility. Scheme programmers can add new functionality on the fly as the program runs; they can do clever things with functions that C programmers can only mimic imperfectly and at much greater length.[45] It is easier to work with and reason about functions in Scheme than in C, because functions are first-class in Scheme and not in C.

People are first-class objects on Twitter: it has the capacity to distinguish and reason about them. The same is true in the many other systems that give people unique identifiers as a way of representing them in databases. Both halves of the phrase are illuminating. On the one hand, people are truly *first-class*: this representation enables useful features that connect directly to these individuals' wants and needs. On the other, people are also *objects*: when these systems represent people, it is often without their knowledge or consent.

I have argued that treating people as first-class objects—representing them with digital identifiers—has significant technical and social consequences. Perhaps it should have legal consequences as well. We should expect the creators of these first-class objects to take care to treat people with the respect and concern the name suggests they deserve.

PROGRAMS 76 (2d ed. 1996); Christopher Strachey, *Fundamental Concepts in Programming Languages*, *in* 13 HIGHER-ORDER & SYMBOLIC COMPUTATION 11, 32–34 (2000).

[44] *First Class*, CUNNINGHAM & CUNNINGHAM, INC., http://www.c2.com/cgi/wiki?FirstClass (last visited Mar. 11, 2011).

[45] *See, e.g.*, DANIEL P. FRIEDMAN ET AL., ESSENTIALS OF PROGRAMMING LANGUAGES 24–25 (1992) ("First class procedures contribute greatly to the expressive power of a language.").