Catherine D´Ignazio reflexiona sobre lo que ocurre cuando una institución recopila datos y menciona que el origen de la palabra "data" viene de "aquello que está dado". Así, piensa, es como muchos se acercan a los datos por primera vez, entendiéndolos como algo neutro: información que estaba ahí, que la institución recopiló y guardó.
En el artículo, D´Ignazio señala que la académica Johanna Drucker propone usar en vez de "data" la palabra"capta", que viene de "lo que se toma". Esto porque, como Drucker explica en su paper ´Graphesis: Visual knowledge production and representation´, data "es considerada información objetiva, mientras que capta es información captada porque se ajusta a las reglas e hipótesis establecidas para cierto experimento".
D´Ignazio dice que pensar en "capta" y no en "data" nos ayuda a recordar que los datos no son nunca neutros, sino que se sitúan en un contexto determinado y son recopilados por un motivo dado. Es importante preguntarse por qué los datos fueron recopilados por una institución en particular, cómo los usa y a quién beneficia o perjudica esa información.

Por qué el contexto es difícil

Entender el contexto de la data (capta) es un gran desafío. Primero, porque los datos los recopilan instituciones para sus fines internos, no para que los usen otros. El texto menciona una referencia a Drew Sullivan, quien dice que "los datos existen para servir a la burocracia y no al periodista". El modo de nombrar, la estructura y la organización de la mayoría de las bases de datos se realizan desde la perspectiva de la institución y no en función de encontrar una historia periodística. D´Ignazio da el ejemplo de cuando sus alumnos pasaron varias semanas tratando de entender la diferencia entre las columnas 'PROD.WASTE (8.1_THRU_8.7)' y '8.8_ONE-TIME_RELEASE' de una base de datos de liberación de químicos tóxicos al medio ambiente producidos por corporaciones.
A veces la falta de contexto o metadata en las bases de datos se debe, entre otras cosas, a la necesidad de recursos para generarla. Pero otras veces, si la institución que recopila la información tiene interés en que cierta información no se haga pública, la falta de metadata, usabilidad y contexto la favorecerán.
D´Ignazio da el ejemplo del registro de la Policía de Boston de su programa de detención y registro FIO (Field Interrogation and Observation). Para este programa, la Policía debe registrar la información de las detenciones de particulares que hace en la calle, los interrogatorios, etc. El año 2014 fue obligada a liberar esta información, luego de una acción legal ganada por la American Civil Liberties Union. Así lo hizo, pero cuando alguien buscaba "detención de registro" en el portal donde se publicaron los datos, no aparecía nada. Era necesario saber -y buscar- por el término que la burocracia policial utiliza en ese programa para nombrar la detención de registro. Así, el contexto puede ayudar a entender por qué hay información que no aparece en una búsqueda.
Otras instituciones pueden publicar la información, pero ser poco claras sobre las limitaciones de esos datos, lo que puede hacer que se malinterpreten seriamente. D´Ignazio da el ejemplo del sitio FiveThirtyEight, que tuvo que retractar un reportaje sobre secuestros de niñas nigerianas que hizo basado en información de GDELT "the Global Database for Events, Language and Tone". La historia hablaba de la incidencia de los secuestros, pero los datos que tomaron de GDELT no eran sobre eventos de secuestros, sino reportes de noticias sobre secuestros. En el libro Data Feminism, que D´Ignazio escribió junto a Lauren Klein, explica que GDELT no describió las limitaciones de su data, presionada para atraer financiamiento para hacer investigación científica de big data.

El detective de contexto de 3 pasos

D´Ignazio dice que un periodista de datos se debe convertir en un “detective del contexto”, que logre conectar información que encuentra en hojas de cálculo y bases de datos, con el ambiente en el cual fueron recopilados esos datos. Para entender la data, el periodista debe entender el quién, qué, cuándo, dónde y cómo de la burocracia de donde vienen esos datos.

En sus clases, D´Ignazio usa el modelo “Detective de contexto de 3 pasos”. Estos pasos se pueden hacer en cualquier orden.

1. Descargue los datos y orientese

Esto implica desglosar lo básico sistemáticamente, para poder hacer buenas preguntas. Puede explorar la data con Excel o Google Spreadsheets. D´Ignazio propone responder preguntas como:

“¿Cuántas observaciones (filas de datos) tienes?

¿Cuántos campos (columnas) tienes?

¿Está claro qué cuenta cada fila? (Recuerde esas incidencias de secuestro versus informes de los medios sobre secuestro; es sumamente importante aclarar lo que registran sus datos).

¿Cuál es el período de tiempo de los datos? Use la función "Ordenar" en cualquier columna con fechas o marcas de tiempo para ver cuándo comienzan y cuándo terminan los datos.

¿Cuál es la extensión geográfica de los datos?

¿Parece que faltan muchos datos?”.

Esta etapa puede ser tan desafiante, que la autora creó una herramienta gratuita online llamada WTFcsv (WTF está pasando con mi archivo .csv).

Lo que hace WTFcsv es analizar cada columna de información y visualizar los datos en términos de patrones por columna. Da el ejemplo de los pasajeros del Titanic. WTFcsv visualizó la información de “sexo” con un gráfico de columnas que muestra que había 314 mujeres y 577 hombres registrados en el Titanic.

Acá lo clave es hacerse buenas preguntas antes de empezar una historia y WTFcsv puede responder todas las preguntas citadas más arriba. D´Ignazio menciona que buenas preguntas sobre esta data serían, por ejemplo, en la calidad de los datos, el saber si la información está completa; en la ética de los datos, el por qué la variable “sexo” es binaria: en el análisis, si los hombres o las mujeres sobrevivieron más; y en el formato, qué significa cierta variable.

2. Explore todos los metadatos disponibles

La metadata es “la data de la data”. En el mundo ideal, todas las bases de datos tendrían un diccionario detallado y actualizado explicando, por ejemplo, qué significan las variables, cuáles son las limitaciones de los datos, las unidades de medida, etc.

A veces encontrar la metadata es difícil porque esta puede ser negada o puede estar desactualizada. Otras veces, puede haber un diccionario de la data pero no llamarse así. D´Ignazio ejemplifica: el diccionario del City of Boston 311 data, se llama ´CRM Value Codex’.

Siempre hay que ser escéptico, chequear toda la información y verificar la data.

3. Investigue los antecedentes del conjunto de datos

Los periodistas normalmente hacemos una investigación de background o antecedentes de nuestras fuentes o de un tema, y hay que hacerlo también en la data. Eso permitirá entender las limitaciones, evitar errores y descubrir historias noticiosas.

Esta etapa del detective de contexto se aplica a tres cosas:

3.1 Investigue los antecedentes de cómo se recopilaron los datos

Heather Krause, consultora de ciencia de datos, usa un documento para crear lo que llama "biografías de datos" donde describe de dónde provienen los datos, quién los recopiló y cómo los recopilaron. Los detalles burocráticos de la producción de los datos es esencial para comprender dónde se pueden introducir errores o datos faltantes (por ejemplo, si los hizo un ser humano, si los midió una máquina, si los reportaron usuarios; o si la forma en que la organización cuenta y mide los datos ha cambiado recientemente afectando la posibilidad de hacer comparaciones).

En este proceso primero miras la metadata. Luego, hablas con una fuente humana. Hay que ser creativos para encontrar quién puede hablar sobre los datos, en caso de que no pueda hablar alguien involucrado en el proceso mismo de recopilación.

Kraus ofrece un modelo para hacer la biografía de los datos que se encuentra linkeado en el texto.

3.2 Investigue los antecedentes de la organización que los recopiló

Es importante conocer qué motivó a una organización a recopilar esos datos y saber cómo los usa.

En el ejemplo de la Policía de Boston, esto implica reportear: "¿Cuál es su misión? ¿Cuánto tiempo han existido? ¿Cuál es su presupuesto? ¿Cuántos oficiales hay? ¿Cuándo han estado en las noticias en los últimos diez años y por qué? También significa investigar el programa FIO específicamente: ¿Cuándo y por qué (la Policía de Boston) comenzó el programa? ¿Era parte de una ola nacional de programas FIO? ¿Existe un debate académico y legal sobre si estos programas son constitucionales y efectivos para reducir la delincuencia?".

Con ese contexto, hay que pensar cómo se usa esa información internamente, por ejemplo en el caso de la Policía de Boston, ¿a quién le reporta esos números? ¿tienen metas o cuotas que cunmplir?, etc.

Las entrevistas con personas del sistema son muy relevantes. Si no las puedes obtener, puedes basarte en el paso siguiente:

3.3 Investigue los antecedentes del entorno regulatorio

Los datos son caros de recopilar, organizar y mantener, y la mayoría de las organizaciones que lo hacen se debe a que están cumpliendo con ciertas leyes o políticas internas. Entender el marco regulatorio arroja luces sobre el porqué una institución recopila ciertos datos, a quién se los reporta y cómo. Por ejemplo, en Estados Unidos, las instituciones de educación superior acreditadas registran e informan agresiones sexuales en los campus porque obedecen las Jeanne Clery Disclosure of Campus Security Policy y la Campus Crime Statistics Act (Clery Act).

Es más difícil reportear los antecedentes de las políticas internas de ciertas organizaciones, pero hay maneras de hacerlo. En los países con leyes de acceso a información pública, por ejemplo, se puede solicitar documentos de gobernanza organizacional y manuales de capacitación para entender el contexto normativo interno que guía la recopilación de los datos.

Trampas

Hay dos trampas que hay que tener en consideración:

1.- Cuidado con hacer suposiciones personales para llenar los vacíos de información. Se debe seguir el consejo de Jonathan Stray "considerar múltiples explicaciones para los mismos datos, en lugar de aceptar la primera explicación que tenga sentido". Por ejemplo, los estudiantes de D´Ignazio estaban analizando

una base de datos de perros y había una variable llamada raza. Uno de los valores era “desconocida” y un alumno interpretó que significaba “raza mixta”, pero en realidad era que no se había llenado ese campo al hacer los registros.
2.- Hay que considerar que hay desequilibrios de poder en el proceso de recolección, la organización y el entorno regulatorio. Así, “los números pueden parecer contar una historia en la primera exploración, pero esa historia puede ser completamente falsa porque el entorno de recolección ha silenciado sistemáticamente a las personas con menos poder”. Fuerzas como el racismo, el patriarcado o el clasismo pueden llevar a contar de menos, o a contar de más, a mujeres y otros grupos marginalizados. Por eso es tan relevante establecer el contexto y no quedarse sólo con el número al pie de la letra. Un ejemplo: la investigación de estudiantes de D´Ignazio reveló que los campuses con mayor número de denuncias de acoso sexual tenían de hecho mejores políticas que aquellos con bajas cifras, y por eso las víctimas se inclinaban en denunciar.

Oportunidades

Hay periodistas y organizaciones de medios creando recursos útiles en base a su reporteo del contexto de datos. Por ejemplo, ProPublica creó Dollars for Docs, que se convirtió en fuente de nuevas investigaciones sobre la influencia de empresas farmacéuticas en contextos locales en Estados Unidos. Así, ProPublica convirtió el contexto de su propia investigación y data en un recurso que pueden usar otras organizaciones.
Los datos verificados y la investigación de contexto pueden ser también una buena fuente de ingresos para los medios. ProPublica tiene a la venta conjuntos de datos sobre diversos temas. Muchos de ellos vienen con una “guía de usuario de datos” - como acuñó Bob Gradeck- algo que trasciende el diccionario de data e incluye cosas como el origen de los datos, cómo los usa la organización y cuáles son sus limitaciones.
La Associated Press (AP) compiló una base de datos nacional sobre segregación escolar en EE. UU. Está a la venta con una guía de usuario de datos de 20 páginas que incluye dónde se recopilaron los datos y qué tipo de preguntas se pueden responder con ellos. También está desarrollando un modelo de suscripción donde las organizaciones pueden pagar por el acceso a conjuntos de datos, su contexto y discusiones con los reporteros que trabajaron en los temas que se generaron en base a esos datos.

Conclusión

Poner los datos en contexto es un trabajo arduo, pero sumamente necesario para hacer periodismo de datos, ya que solo podemos comprender la historia bien, si comprendemos el contexto de los datos.