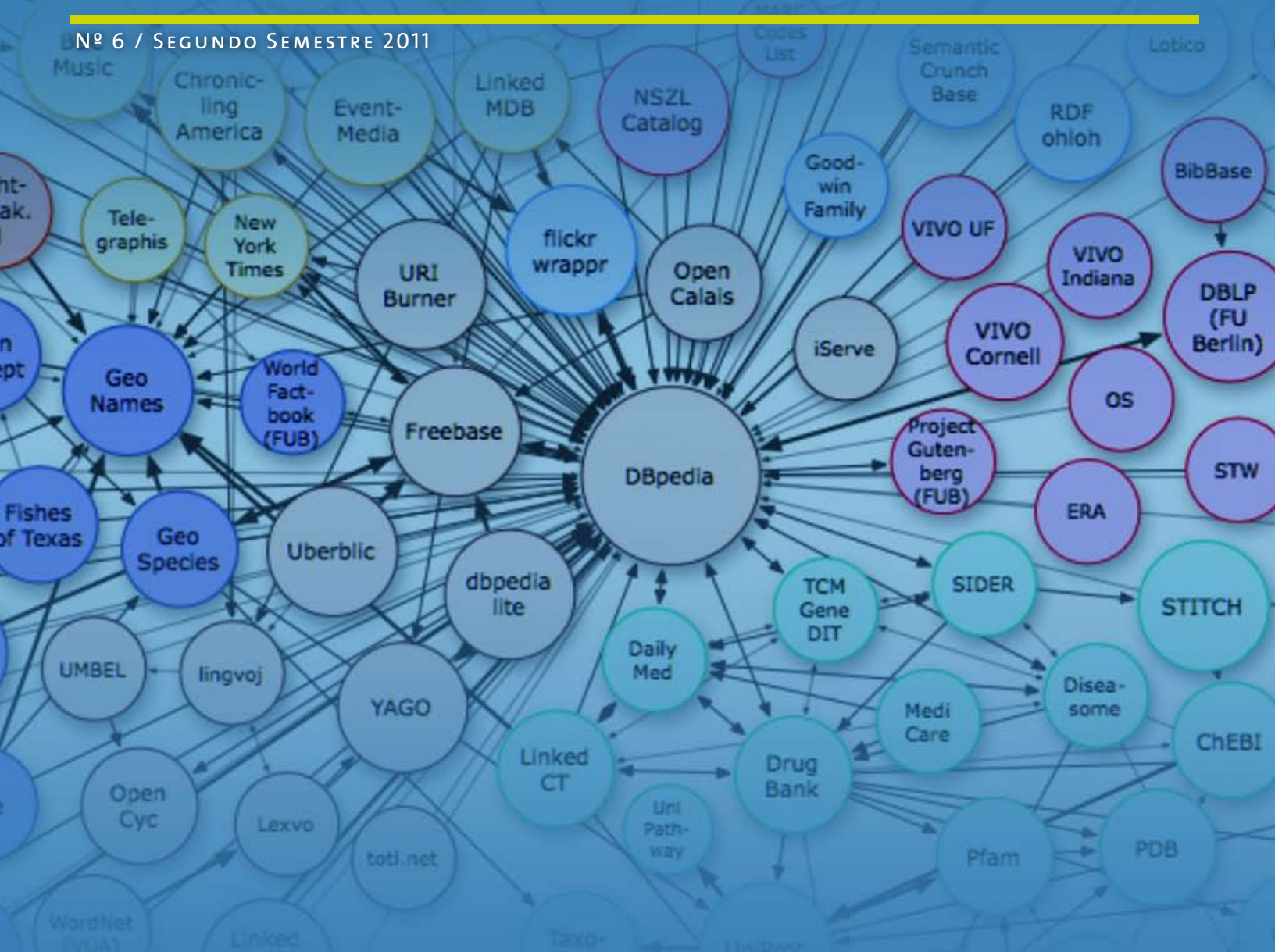
REVISTA
BITS de Ciencia
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

Nº 6 / Segundo Semestre 2011



Alejandro Barros:

Open Data: nuevo paradigma en el manejo de datos

- **En camino hacia la Web Semántica:** Experiencias de la Biblioteca del Congreso Nacional de Chile
- **Ricardo Baeza Yates:** 32 años de computación: de estudiante a Fellow

REVISTA
BITS de Ciencia
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

Dirección
Departamento de Ciencias de la Computación
Avda. Blanco Encalada 2120, 3º piso
Santiago, Chile.
837-0459 Santiago
www.dcc.uchile.cl
Teléfono: 56-2-9780652
Fax: 56-2-6895531
revista@dcc.uchile.cl

Revista Bits de Ciencia N°6
ISSN 0718-8005 (versión impresa)

www.dcc.uchile.cl/revista
ISSN 0717-8013 (versión en línea)

CONTENIDOS

EDITORIAL

La Computación está más viva que nunca. Todo el espectro de nuestros campos - incluyendo la teoría, la simulación, la implementación, etc.- debería estar más atento y ser más propositivo que nunca antes en la historia. Esto porque la sociedad del futuro (que ya es la sociedad del presente en un puñado de países), la tan manoseada pero no por eso menos importante sociedad del conocimiento, solo podrá ser construida a través de la aplicación masiva de las tecnologías de la información.

Uno de los mayores desafíos que esta sociedad moderna impone en nuestra área es la inmensa proliferación de datos de todo tipo. Cada vez producimos más información, y cada vez tenemos mayor capacidad computacional para almacenarla. Por ejemplo, las bases de datos astronómicas generan diariamente terabytes de información describiendo el estado del cielo, las bases de datos genómicas describen secuencias de ADN de muchísimos organismos, nuestro paso por las redes sociales genera cada vez más datos sobre nuestros gustos, relaciones y posición geográfica, entre muchos otros. Lo importante de todo esto, es notar que el real valor de la sociedad del conocimiento está en esos datos. O puesto de otra forma, los datos serán en la sociedad del conocimiento lo que alguna vez fue el oro o el dinero.

Este paso hacia una sociedad basada en la información ha hecho cambiar muchos de nuestros paradigmas. Por ejemplo, Tim Berners-Lee –el creador de nuestro fetiche moderno más relevante, la Web– declaró hace poco que ésta debía pasar lo antes posible de su actual estado centrado en documentos (es decir, donde los documentos o páginas son lo más importante) a un estado centrado en datos (es decir, donde los datos sean ciudadanos de primera clase). El modelo de esta Web del futuro puede verse como el de una inmensa red de bases de datos distribuidas, que colaboran activamente intercambiando su información. Este es el famoso concepto de Linked Data.

Pero si los datos son el capital del futuro, entonces, ¿quiénes deberían ser los dueños de esos datos? Esta es sin duda una decisión política. Pero lo que es claro es que si queremos que la sociedad del conocimiento sea, a la vez, la sociedad de la inclusión, entonces una respuesta justa a esta pregunta sería: "Todos". Todos deberíamos ser dueños de la mayor cantidad de datos posibles (con las debidas reservas de privacidad). Lo más interesante de esto es que la información se presta perfectamente para esta idea: al contrario del capital que es limitado, los datos claramente abundan; los datos no se gastan, y además pueden ser replicados y compartidos.

Pero todo esto que suena tan bonito es bastante más complejo en la práctica. Por ejemplo, por el momento muchas de las compañías más exitosas se están haciendo gratuitamente con nuestros datos (supongo que no necesito nombrarlas). Además, en general los datos son como el material en estado crudo, no refinado. Saber sacar la información relevante que hay en ellos es una habilidad que debe ser enseñada y entrenada. Es como si a uno le regalaran un cerro que está lleno de oro. Para hacerse rico hay que saber sacarlo.

Pero no nos adelantemos. El tema de la Revista es Open Data, es decir, la idea de hacer públicos la mayor cantidad de datos posibles a la mayor cantidad de gente posible. Esto acarrea problemas que van desde los legales, hasta los más técnicos que tienen que ver con el formato de publicación de esos datos. Lo que hemos tratado de hacer en este número de la Revista es acercarnos a esos problemas a través de la visión de varios expertos: Alejandro Barros, sobre el paradigma de Open Data; Álvaro Graves, sobre la aplicación de este paradigma en los gobiernos; la Biblioteca del Congreso Nacional, acerca de cómo esta entidad está dejando disponibles sus datos; Jens Harding, sobre el concepto de open software; Julio Costa, sobre un sistema abierto de mapas; Karim Pichara, Rodolfo Angeloni y Susana Eyheramendy, acerca de Datos Astronómicos; y Alejandro Hevia, sobre temas de privacidad.

Por otro lado, también seguimos con nuestras secciones habituales: Investigación Destacada, Computación y Sociedad (con artículos de Jorge Olivos, Ricardo Baeza Yates y Juan Álvarez), Surveys (donde Claudio Gutiérrez y Daniel Hernández nos cuentan sobre la Web de Datos), y Conversaciones (con el destacado profesor Héctor García-Molina, de la Universidad de Stanford).

¡Esperamos les guste!

Pablo Barceló
Editor Revista Bits de Ciencia

Expresiones regulares (autómatas) **con variables y sus aplicaciones**

En este artículo estudiamos expresiones regulares que utilizan tanto símbolos de un alfabeto finito como variables. Tales variables se interpretan como símbolos en el alfabeto. Además, consideramos dos tipos de lenguajes definidos por estas expresiones: bajo la semántica *existencial,* una palabra pertenece al lenguaje de la expresión con variables E si pertenece al lenguaje definido por alguna expresión que se puede obtener desde E al reemplazar variables por símbolos; bajo la semántica *universal,* una palabra pertenece al lenguaje de la expresión con variables E si pertenece al lenguaje definido por toda expresión que se puede obtener desde E al reemplazar variables por símbolos. Tales lenguajes son regulares, y además demostramos que aparecen naturalmente en varias aplicaciones como consultar bases de datos de grafos con información incompleta y el análisis de programas. Para proveer un análisis computacional más sólido, mencionamos también ciertos resultados teóricos que ayudan a entender el comportamiento de las expresiones regulares con variables, así como la complejidad de algunos de los problemas de decisión más básicos asociados con ellas.

Organización: en la siguiente sección introducimos las definiciones básicas de lenguajes regulares y autómatas que son

Pablo Barceló
Profesor Asistente DCC, Universidad de Chile. Ph.D. in Computer Science, University of Toronto (2006); Magíster en Ciencias de la Computación, Pontificia Universidad Católica de Chile (2002); Ingeniero en Electricidad, Pontificia Universidad Católica de Chile. Áreas de interés: Bases de Datos, Lógica para la Ciencia de la Computación, autómata.
pbarcelo@dcc.uchile.cl

necesarias para entender la investigación realizada. Luego, motivamos la introducción de las expresiones regulares con variables con dos aplicaciones diferentes: análisis de programas y bases de datos de grafos con información incompleta. A partir de éstas definimos las dos semánticas para las expresiones regulares con variables: Existencial y Universal. En la siguiente sección listamos y explicamos las propiedades computacionales más importantes de estas expresiones, y finalmente discutimos las conclusiones de nuestro trabajo, así como los problemas relacionados que deseamos estudiar a futuro.

EXPRESIONES REGULARES Y AUTÓMATAS

Las *expresiones regulares* son un método gramatical ampliamente utilizado en computación para especificar conjuntos (posiblemente infinitos) de palabras sobre un alfabeto finito. Su objetivo básico es servir como una herramienta concisa y flexible para la especificación de patrones de texto.

De forma de hacer más fácil la presentación, a partir de ahora reducimos nuestro estudio al alfabeto binario A = {0,1}. Sin embargo, cabe notar que todo lo que mencionemos acerca de las expresiones regulares, y sus variantes, a lo largo del artículo es independiente de esta elección inicial. Es decir, todas las propiedades que mencionemos son también ciertas si trabajamos con cualquier alfabeto A′ distinto de A.

El conjunto de expresiones regulares se halla definido recursivamente por la siguiente gramática:

$$\varphi, \varphi' := \varnothing \mid \varepsilon \mid 0 \mid 1 \mid \varphi \cup \varphi' \mid \varphi \cdot \varphi' \mid \varphi*$$

Para facilitar la simplicidad de notación, el símbolo · comúnmente se elimina de las expresiones regulares (ya que es fácilmente distinguible dentro del contexto, los lugares en que aparece).

Como mencionamos anteriormente, cada expresión regular φ define un conjunto de palabras (es decir, un *lenguaje*) L(φ) sobre el alfabeto A. Tal lenguaje se define recursivamente como sigue:

Casos base:

1) $L(\varnothing) = \varnothing$. Es decir, ∅ denota al conjunto vacío.
2) $L(\varepsilon) = \{\}$. Esto es, ε denota a la palabra vacía.
3) $L(0) = \{0\}$ y $L(1) = \{1\}$. Esto es, el 0 denota a la palabra con un único símbolo 0, y análogamente para el 1.

Casos inductivos:

4) $L(\varphi \cup \varphi') = L(\varphi) \cup L(\varphi')$. Es decir, como era de esperar el símbolo ∪ representa la unión de lenguajes.
5) $L(\varphi \cdot \varphi') = L(\varphi) \cdot L(\varphi')$, donde

   $L(\varphi) \cdot L(\varphi') = \{ww' \mid w \in L(\varphi) \text{ y } w' \in L(\varphi')\}$.

Esto quiere decir que el símbolo · representa la *concatenación* de lenguajes.

6) $L(\varphi*)$ se define como la unión de todos los lenguajes $L(\varphi)^i$, para i mayor o igual a 0, donde los lenguajes $L(\varphi)^i$ se definen inductivamente de la siguiente forma:

   $L(\varphi)^0 := \{\}$ y $L(\varphi)^{i+1} = L(\varphi)^i \cdot L(\varphi)$.

Esto es, el símbolo *, que se llama usualmente *clausura de Kleene*, representa la *concatenación* de un número arbitrario de veces de un lenguaje consigo mismo.

Por ejemplo, la expresión regular (01)* representa el lenguaje de todas las palabras de la forma 0101 ... 0101, mientras 0* ∪ 0*10* ∪ 0*10*10* representa el lenguaje de las palabras con *a lo más dos 1s*. Por supuesto, no todo conjunto de palabras puede ser representado por una expresión regular (es decir, no todo lenguaje es regular). Uno de los ejemplos más paradigmáticos es el lenguaje que contiene a todas las palabras de la forma $0^n 1^n$, para $n > 0$. Esto es, el lenguaje de todas las palabras que comienzan con una secuencia de 0s y luego siguen con una secuencia de 1s del mismo largo. Intuitivamente, este lenguaje no es regular porque las expresiones regulares carecen de la habilidad de "contar", es decir de especificar que hay la misma cantidad de 0s que de 1s en una palabra.

Por razones obvias, el conjunto de los lenguajes definidos por expresiones regulares se llaman *lenguajes regulares*. Aunque es probable que los lenguajes regulares hayan sido creados (¿o descubiertos?) por varios investigadores al mismo tiempo, usualmente se cita como su inventor, en los años cincuenta, al lógico estadounidense Stephen Kleene. También fueron estudiados por el famoso lingüista Noah Chomsky, quien los enmarcó en su jerarquía de gramáticas formales, siendo los lenguajes regulares uno de sus escalones más bajos.

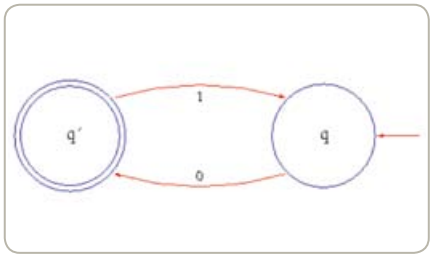
Una de las propiedades más importantes de los lenguajes regulares es su clausura con respecto a las combinaciones Booleanas usuales. Esto quiere decir que los lenguajes regulares son cerrados bajo unión, intersección y complemento. Note que la primera operación (unión) pertenece explícitamente a la gramática, y, por tanto, demostrar que los lenguajes regulares son cerrados bajo unión es trivial. Por otro lado, ni la intersección ni el complemento son operaciones en la gramática, y de hecho probar que los lenguajes regulares son cerrados bajo estas operaciones requiere una demostración no trivial.

Otra buena propiedad de los lenguajes regulares es su "robustez", en el sentido que pueden ser caracterizados de muchas formas diferentes, e.g. como gramáticas, algebraicamente, en términos del poder expresivo de ciertas lógicas, etc. Probablemente, junto con la caracterización gramatical de los lenguajes regulares en términos de expresiones regulares, que ya hemos visto, la otra caracterización

más famosa es la algebraica, en términos de *autómatas*. Esta equivalencia es particularmente importante, ya que en la práctica muchas veces es conveniente construir el autómata equivalente a una expresión regular para poder determinar las palabras que la satisfacen.

Recordemos que un autómata es una tupla $T = (Q,q,F,\delta)$, donde (1) Q es un conjunto finito de *estados,* (2) q es un estado particular en Q denominado estado *inicial,* (3) F es el subconjunto de Q que contiene a los estados *finales,* y (4) δ es una función parcial de *transición,* que se define desde $Q \times \{0,1\}$ en Q. El autómata T *acepta* una palabra $w = a_1\, a_2 \cdots a_n$ sobre alfabeto {0,1} si el autómata puede "correr" sobre w desde el estado inicial a uno final siguiendo las reglas dictadas por la función de transición. Formalmente, si existe función $\rho : \{0,1,\ldots,n\} \rightarrow Q$ tal que (a) $\rho(0) = q$, (b) $\rho(i+1) = \delta(\rho(i),a_{i+1})$, para todo i entre 0 y (n-1), y (c) $\rho(n) \in F$. El lenguaje aceptado por el autómata T se define como el conjunto de todas las palabras aceptadas por T.

Por ejemplo, el siguiente autómata acepta exactamente el lenguaje regular 0(10)*:



Los estados son los círculos azules y las transiciones son las flechas en rojo. El estado inicial es q, denotado por la flecha de entrada sin etiqueta. El estado final es q', denotado por el doble círculo. Desde el estado q la única posible transición es al estado q', pero sólo si el autómata está leyendo el símbolo 0 en la palabra. De la misma forma, desde el estado q' la única posible transición es al estado q, pero sólo si el autómata está leyendo el símbolo 1 en la palabra.

El hecho que un autómata reconozca un lenguaje que es regular no es una coincidencia, ya que el famoso teorema de Kleene establece la equivalencia de los lenguajes regulares y aquellos aceptados por autómatas.

> ***Teorema de Kleene:*** *sea L un lenguaje. Entonces las siguientes afirmaciones son equivalentes:*
>
> *1) L es regular (es decir, L = L(φ) para alguna expresión regular φ).*
>
> *2) L es aceptado por algún autómata T.*

Es interesante ver que, en cierto sentido, el Teorema de Kleene expresa que los lenguajes regulares pueden ser entendidos tanto *declarativamente* - aquellos que son *especificados* por las expresiones regulares - como *proceduralmente* - aquellos que son *aceptados* por los autómatas. Este tipo de equivalencias aparecen en distintas áreas de la computación y son comúnmente muy relevantes. Por un lado, la parte declarativa permite al usuario especificar lo que desea, mientras la parte procedural ayuda en la implementación y optimización del sistema. Un ejemplo paradigmático es el de los lenguajes de consulta para bases de datos relacionales: el usuario especifica sus consultas en SQL, mientras el sistema las ejecuta utilizando el álgebra relacional.

Es importante destacar que las expresiones regulares y su contraparte algebraica, los autómatas, tienen variadas aplicaciones en distintas áreas de la computación. Éstas incluyen, entre innumerables otras, lingüística computacional [Mar05], compiladores [ALSU06], búsqueda en texto [CR03, NR07], datos semiestructurados [Bun97, BSV99], verificación de software [VW86], análisis de programas [NNH10], etc.

EXPRESIONES REGULARES CON VARIABLES

Nuestra investigación trata sobre el tema de las Expresiones Regulares con Variables (ERVs), las que permiten describir sucintamente otras expresiones regulares más complejas. Para motivar el problema comenzaremos con una aplicación de estas expresiones en el análisis de programas [LRYS04].

Utilizamos el alfabeto que contiene las operaciones sobre los elementos del programa, e.g. variables, punteros, archivos, etc. (por ejemplo, **def**, para definir una variable, **use**, para usarla, **malloc**, para localizar un puntero, entre otros). A estos símbolos comúnmente sigue una variable; e.g. **def**(x) significa definir la variable x.

En este caso las ERVs sirven para describir y detectar cierto tipo de *bugs* (i.e. comportamiento indeseado) en el programa. Por ejemplo, la expresión:

$$(\neg \textbf{def}(x))^* \ \textbf{use}(x),$$

donde $\neg$**def** es una abreviación para el complemento de la expresión regular def, que identifica aquellas variables que han sido utilizadas sin antes ser definidas. En general, dada una ERV E y un programa P, nos interesa encontrar los posibles valores de las variables que hacen que E se satisfaga en P (estos valores corresponden a los *bugs* de P). En términos más formales, nos interesan aquellas palabras que son definidas por *alguna* expresión regular sin variables que se pueda obtener desde E reemplazando a las variables por símbolos del alfabeto. Es decir, en este caso interpretamos existencialmente la semántica de la expresión regular con variables.

Una segunda aplicación de las ERVs viene del área de *Bases de Datos de Grafos* (GDBs). En su forma más simple una GDB es un par (V,E), donde V es un conjunto finito de

vértices y E es un subconjunto de triples en V × A × V, donde A es un alfabeto finito. Es decir, E es un conjunto de arcos etiquetados en A. Por ejemplo, sea C el conjunto de todas las ciudades del mundo y F el conjunto de todos los triples de la forma (c,a,c′), tal que c y c′ son ciudades en C y a es el nombre de una aerolínea que vuela en forma directa desde c hasta c′. Entonces H = (C,F) es una GDB.

El lector familiarizado con los temas de autómatas notará que una GDB no es más que un autómata *no determinista*, donde los vértices corresponden a los estados y las arcos a las transiciones. Como es bien sabido, estos autómatas no entregan mayor poder expresivo a su versión determinista (la que fue definida en la sección anterior). Es decir, todo lenguaje definido por un autómata no determinista puede también ser definido por un autómata, y por tanto también por una expresión regular. Esto quiere decir que las GDBs son, en esencia, expresiones regulares.

Como es usual en cualquier modelo de base de datos, las GDBs vienen acompañadas de su propio lenguaje de consulta, el que permite extraer información a partir de la información almacenada. Una de las características más importantes de las GDBs, y que en particular la hacen distinas de las bases de datos relacionales, es que en ellas estamos tan interesados en consultar la "topología" de los datos como los datos mismos. En particular, muchas veces estamos interesados en "navegar" el grafo, es decir, en recorrerlo recursivamente desde un lado a otro. Una típica consulta de esta forma es verificar si un par de nodos se haya unido por una palabra en un lenguaje regular dado. Por ejemplo, continuando con la GDB H = (C,F), uno podría querer verificar si existen vuelos (no necesariamente directos) entre Santiago y Toronto que sólo utilicen a LAN o Air Canada. Para ello se debe verificar si existe un camino desde Santiago a Toronto en H tal que la concatenación de las etiquetas en ese camino pertenece al lenguaje regular definido por (LAN ∪ Air Canada)*.

> Dado que la mayoría de los problemas de decisión clásicos de autómatas y expresiones regulares se vuelven intratables en la presencia de variables, es importante en el futuro desarrollar heurísticas que permitan trabajar con ellas en ciertos contextos importantes.

En la mayoría de las aplicaciones modernas, en que los datos son constantemente transferidos, intercambiados y tienen niveles no menores de incertidumbre, es necesario proveer un modelo flexible que permita especificar que ciertos datos son desconocidos o simplemente no están disponibles. Este modelo se denomina de información *incompleta*. Usualmente en este escenario buscamos las respuestas a las consultas que son independientes de la interpretación faltante. Tales respuestas se denominan *certeras*, ya que son invariantes frente a la reparación que establezcamos sobre la base de datos.

En un reciente artículo [ABR11] hemos comenzado el estudio de este tema sobre GDBs. Una de las mayores fuentes de incompletitud en este escenario es la pérdida de información estructural, principalmente la pérdida de la información contenida en la etiqueta de un arco. Como ya mencionamos, cada GDB puede ser descrita por un autómata, y por tanto por una expresión regular. Eso quiere decir que las GDBs con información estructural incompleta pueden ser descritas por expresiones regulares con variables (donde las variables representan la información estructural faltante).

Recuerde que en la presencia de información incompleta estamos interesados en las respuestas certeras a una consulta. En particular, pensando en el lenguaje de consulta específico de las GDBs, quisiéramos detectar si dos nodos se hayan unidos por una palabra en un lenguaje regular dado, independiente de la interpretación de la información faltante (las variables). Esto quiere decir que, al contrario del caso anterior, en este escenario nos interesa una interpretación *universal* de las ERVs que representan nuestras GDBs con información parcial.

Podemos entonces definir la semántica de una ERV E. Primero debemos entender cómo interpretar sus variables. Para ello ocupamos el concepto de valuación, que no es más que una función $\eta : V \to A$, donde V es el conjunto de variables mencionadas en E. Esto es, η es una función que a cada variable le asigna una letra del alfabeto. Definimos además la aplicación de η sobre E, denotado $\eta(E)$, como la expresión regular *sin variables* que se obtiene desde E al reemplazar simultáneamente cada variable x en E por el símbolo $\eta(x)$.

Dada una expresión regular E con variables, definimos su semántica tanto en términos

existenciales como universales (motivadas por las aplicaciones mencionadas más arriba) como sigue:

a) La semántica existencial de E es el lenguaje $L_{\cup}(E)$ que se define como la unión de todos los lenguajes de la forma $L(\eta(E))$, donde η es una valuación de E. Esto es, una palabra pertenece a $L_{\cup}(E)$ si y sólo si es aceptada por alguna expresión regular E' sin variables que se puede obtener desde E al reemplazar simultáneamente a sus variables por símbolos del alfabeto.

b) La semántica universal de E es el lenguaje $L_{\cap}(E)$ que se define como la intersección de todos los lenguajes de la forma $L(\eta(E))$, donde η es una valuación de E. Esto es, una palabra pertenece a $L_{\cap}(E)$ si y sólo si es aceptada por toda expresión regular E' sin variables que se puede obtener desde E al reemplazar simultáneamente a sus variables por símbolos del alfabeto.

Por ejemplo, sea E la expresión $(0 \cup 1)^*$ xy $(0 \cup 1)^*$. Entonces, la palabra 00 pertenece a $L_{\cup}(E)$, como lo atestigua la valuación $\eta(x) = \eta(y) = 0$. Por otro lado, la palabra 10011 pertenece a $L_{\cap}(E)$ y no hay palabra de largo menor que también pertenezca a este lenguaje. Esto se debe al hecho que xy es una subexpresión de E, y que, por tanto, si una palabra w pertenece a $L_{\cap}(E)$ entonces w contiene a toda palabra de largo 2 como subpalabra. Es posible demostrar por una simple enumeración de casos que no hay palabra más corta que 10011 que tenga esta propiedad (y note que 10011 sí la tiene).

Note que cada ERV E utiliza un número finito de variable y que, por tanto, el número de posibles valuaciones para E es también finito. Esto nos permite hacer una importante primera observación: dado que los lenguajes regulares son cerrados bajo unión e intersección, entonces tanto $L_{\cup}(E)$ como $L_{\cap}(E)$ son también lenguajes regulares. Esto quiere decir que las ERVs no agregan poder expresivo a las expresiones regulares sin variables. Pero como veremos en la próxima sección, lo que sí aportan es la capacidad de expresar algunos lenguajes de forma más sucinta.

Es importante hacer notar que ésta no es la única semántica posible para las expresiones regulares con variables. De hecho, hace varias décadas la comunidad de lenguajes formales viene estudiando la clase de los *patrones* [Sal03], que son nada más que concatenaciones de letras y variables. En el caso de los patrones, sin embargo, las valuaciones permiten reemplazar variables no sólo por letras sino también por palabras arbitrarias en el alfabeto (bajo una semántica existencial). Esto hace que los patrones, al contrario de las ERVs, puedan expresar propiedades mucho más allá del mundo regular. Por ejemplo, el patrón xx define el conjunto de todas aquellas palabras de la forma ww, donde w es una palabra cualquiera. Este lenguaje no es regular y, de hecho, ni siquiera es *context-free*. Este aumento en expresividad conlleva, como es de esperar, problemas en términos de la decibilidad de algunos problemas básicos de decisión (por ejemplo, la equivalencia) [JSSY05], que sí son decidibles para las expresiones regulares (y, por tanto también para las ERVs, dado que éstas no pueden expresar lenguajes no regulares).

PROPIEDADES COMPUTACIONALES DE LAS ERVs

Aunque las ERVs constituyen un modelo simple de especificación de propiedades importantes en distintas áreas de la computación, hasta el momento de nuestro trabajo nadie había iniciado un estudio sistemático de sus propiedades computacionales. A continuación detallamos algunas de las más importantes conclusiones obtenidas a lo largo de nuestro estudio [ABR11a]:

1) Hemos mencionado en la sección anterior que las ERVs sólo pueden definir, bajo ambas semánticas, lenguajes regulares. Esto se debe al hecho que los lenguajes regulares son cerrados bajo unión e intersección, y la semántica existencial y universal de una ERV corresponden a la unión e intersección, respectivamente, de todas las ERVs sin variables obtenidas desde E al reemplazar sus variables por letras en el alfabeto (i.e. por sus valuaciones). Sin embargo, el número de valuaciones de E es exponencial en el número de sus variables, y, por tanto, $L_{\cup}(E)$ y $L_{\cap}(E)$ pueden ser expresados, respectivamente, por expresiones regulares de tamaño exponencial, en el caso de la semántica existencial, y doble exponencial, en el caso de la semántica universal.

Esto sugiere fuertemente que las ERVs son al menos exponencialmente más sucintas que su versión sin variables. Es decir, que ERVs de tamaño polinomial pueden definir lenguajes que sólo pueden ser definidos por expresiones regulares de tamaño exponencial. Note que esto no se sigue directamente de lo explicado en el párrafo anterior, ya que es necesario demostrar que existen ERVs E de tamaño polinomial tal que $L_{\cup}(E)$ (o $L_{\cap}(E)$) *requieren* para ser expresados de expresiones regulares de tamaño exponencial. Éste es uno de los principales resultados de nuestro artículo:

> ***Teorema:*** *existe una familia* $\{En\}_{n>0}$ *de ERVs de tamaño polinomial en n, tal que cualquier:*
>
> *a) expresión regular, o*
>
> *b) autómata (determinista o no determinista),*
>
> *que define* $L_{\cup}(E_n)$ *(respectivamente,* $L_{\cap}(E_n)$*) es de tamaño al menos exponencial (respectivamente, doble exponencial) con respecto a* E_n.

Esto significa que efectivamente las ERVs son al menos exponencialmente más sucintas que su versión sin variables, y que este resultado es robusto, en el sentido que aplica a todos los modelos de computación equivalentes a las expresiones regulares que hemos descrito en el presente artículo.

2) Existe otra interesante forma de demostrar que las ERVs permiten describir sucintamente lenguajes regulares complejos. Un problema clásico en lenguajes regulares es el de intersección, definido de la siguiente forma:

dadas expresiones regulares E_1 , ... , E_k de tamaño O(n), determine el tamaño de la menor expresión E que es equivalente a la intersección de todos los E_i´s (tal E existe ya que las expresiones regulares son cerradas bajo intersección). Es bien sabido que en algunos casos el menor E que cumple la propiedad expresada arriba es de tamaño $O(n^k)$ (es decir, exponencial). Sin embargo, utilizando ERVs y la semántica universal podemos expresar la intersección polinomialmente:

Teorema: *Sean E_1 , ... , E_k expresiones regulares de tamaño O(n). Existe ERV E de tamaño O(nk) tal que $L_\cap(E)$ es equivalente a la intersección de todos los E_i´s.*

Por supuesto, todo este poder de concisión de las ERVs tiene costos al momento de analizar la complejidad de ciertas tareas básicas de las expresiones regulares, lo que veremos en el próximo punto.

3) El problema básico de decisión para una expresión regular es el de verificar si una palabra pertenece al lenguaje definido por ella. Esto es, dada expresión regular φ y palabra w, ¿es cierto que $w \in L(\varphi)$? Es fácil ver que este problema puede ser resuelto eficientemente. Una demostración usual procede como sigue: convierta φ en un autómata no determinista equivalente. Se sabe que esto se puede hacer en tiempo polinomial. Luego verifique si la palabra w es aceptada por el autómata, lo que también puede realizarse polinomialmente.

Para el caso de las ERVs esto no es tan fácil. Imagine que dada ERV E queremos verificar si una palabra w pertenece a $L_\cup(E)$ o $L_\cap(E)$. La técnica de primero convertir a $L_\cup(E)$ o $L_\cap(E)$ en un autómata equivalente T, para luego verificar si w es aceptado por T, nos acarrearía costos computacionales insalvables, ya que sabemos que T puede ser de tamaño exponencial en E. Es decir este procedimiento nos entrega un algoritmo al menos exponencial, y, por tanto, no implementable computacionalmente.

Es posible demostrar que la complejidad del problema es un poco mejor, aunque no mucho más, utilizando el concepto de valuaciones. Note que para verificar que w pertenece a a $L_\cup(E)$ basta "adivinar" una valuación η para E y verificar, en tiempo polinomial, que w está en a $L(\eta(E))$. Dado que η es un testigo de tamaño polinomial en E, podemos concluir que el problema puede ser resuelto en *nondeterministic polynomial time* (NP). De la misma forma, verificar si w está en $L_\cap(E)$ puede ser resuelto en el complemento, coNP, de la clase NP.

Lamentablemente, salvo en casos bastante restringidos, el problema de verificar si una palabra w está en $L_\cup(E)$ puede ser completo para la clase NP. Esto dice que, bajo suposiciones ampliamente diseminadas en la comunidad, el problema es computacionalmente intratable. En particular, la complejidad del problema coincide con la complejidad del problema de satisfacibilidad o del vendedor viajero, ambos considerados inherentemente exponenciales.

CONCLUSIONES

Hemos definido la clase de las ERVs, analizado su aplicabilidad en dos áreas distintas de la computación y estudiado algunas de sus propiedades computacionales básicas. Dado que la mayoría de los problemas de decisión clásicos de autómatas y expresiones regulares se vuelven intratables en la presencia de variables, es importante en el futuro desarrollar heurísticas que permitan trabajar con ellas en ciertos contextos importantes (e.g. Bases de Datos de Grafos).

Otro problema importante es el de la clausura de las ERVs con respecto a las combinaciones Booleanas. Por ejemplo, sabemos que si E y E′ son ERVs entonces $L_\cup(E) \cap L_\cup(E')$ sólo puede ser representado, en algunos casos, por una expresión regular sin variables de tamaño al menos exponencial en E y E´. Sin embargo, es aún posible que el mismo lenguaje pueda ser representado por una ERV de tamaño polinomial. Este tipo de resultados podrían ser útiles en contextos dinámicos en que las ERVs son permanentemente actualizadas, modificadas y consultadas, como es el caso de las Bases de Datos de Grafos con información incompleta.

AGRADECIMIENTOS

Este trabajo fue realizado en colaboración con Leonid Libkin y Juan Reutter, ambos de la Universidad de Edinburgo. Mi trabajo fue patrocinado por el proyecto Fondecyt 1110171. BITS

BIBLIOGRAFÍA

[1] [ALSU06] A. Aho, M. Lam, R. Sethi, J. Ullman. Compilers: Principles, Techniques and Design. Addison-Wesley, 2nd edition, 2006.

[2] [BLR11] P. Barceló, L. Libkin, J. Reutter. Querying Graph Patterns. PODS 2011.

[3] [BLR11a] P. Barceló, L. Libkin, J. Reutter. Parameterized Regular Expressions and Their Languages. Enviado.

[4] [Bun97] P. Buneman. Semistructured Data. PODS 1997.

[5] [BSV99] P. Buneman, D. Suciu, V. Vianu. Data on the Web. Morgan Kauffman, 1999.

[6] [CR03] M. Crochemore, W. Rytter. Jewels of Stringology. WSP, 2003.

[7] [JSSY95] T. Jiang, A. Salomaa, K. Salomaa, S. Yu. Decision problems for patterns. JCSS, 50(1), 53-63, 1995.

[8] [LRYS04] Y. Liu, T. Rothamel, F. Yu, S. Stoller, N. Hu. Parametric Regular Path Queries. PLDI 2004.

[9] [Mar05] C. Martín-Vide. Formal Grammars and Languages. The Oxford Handbook of Computational Linguistics, 2005.

[10] [NR07] G. Navarro, M. Raffinot. Flexible Pattern Matching in Strings: Practical On-Line Search Algorithms for Texts and Biological Sequences. Cambridge University Press, 2002.

[11] [NNH10] F. Nielson, H. Nielson, C. Hankin. Principles of Program Analysis. Springer-Verlag, 2010.

[12] [Sal03] K. Salomaa. Patterns. Bulletin of the EATCS, 2003.

[13] [VW86] M. Vardi, P. Wolper. An Automata-Theoretic Approach to Automatic Program Verification. LICS 1986.

El DCC 1983 – 1988: **refundando el Departamento**

Jorge Olivos (a la izquierda) durante una presentación de la empresa Epson. **Gentileza: Gastón Carreño.**

Jorge Olivos
Académico Departamento de Ingeniería Matemática, Universidad de Chile (1970-1982). Director DCC Universidad de Chile (1983 – 1988). Actualmente, Director del Centro de Computación, FCFM, Universidad de Chile.
jolivos@dcc.uchile.cl

En dos artículos previos[1], se relata el nacimiento, en 1975, del Departamento de Ciencias de la Computación (DCC) de la Universidad de Chile. Sin embargo, en 1981 dejó de serlo a raíz de una decisión administrativa. Fue necesaria una década para recuperar en pleno derecho el estatus de Departamento. Como Director desde 1983 a 1988 estuve a cargo de impulsar las gestiones para devolverle la categoría perdida a la unidad. En el presente artículo relato esta experiencia -no exenta de dificultades- de cómo aporté a sentar las bases de lo que hoy tenemos.

EL CONTEXTO DE UNA DECISIÓN

Cuando en 1983 acepté dirigir la entonces División de Computación, nunca imaginé las dificultades que tendría que sortear. Mirando hacia atrás, creo que esta historia tiene innumerables aristas, la mayoría marcadas por trabas y conflictos. Sin embargo, pienso que, sobre todo, éste fue un período que se caracterizó por el gran esfuerzo realizado por llevar adelante este proyecto.

En una reunión con Pepe Pino, Alfredo Piquer y Patricio Poblete se me solicitó hacerme cargo de esta unidad. Conocía muy bien a Pepe, ya que habíamos sido compañeros de carrera. A Alfredo lo había tenido como alumno destacado en la carrera de Ingeniería Matemática y como, además, tenía buenos antecedentes de Patricio, finalmente acepté el desafío.

Para ese entonces, la disciplina había cobrado importancia en todas las universidades del mundo desarrollado, por lo que me parecía razonable que se pudiese reproducir en Chile lo que ya era una realidad en otros lugares. La misión, entre muchas otras, era volver a formalizar en una estructura departamental

1 Revista Bits de Ciencia números cuatro y cinco.

un cuerpo de conocimientos donde la Facultad ya tenía un claro liderazgo a nivel país gracias, en particular, a las personas que formaban la División.

En la primera reunión me quedó claro que debía aceptar la Dirección, a fin de evitar la disolución definitiva del proyecto. La situación de la unidad era compleja. Las dificultades crecientes que hubo los años anteriores, habían hecho mella en el grupo.

Ese fue el escenario con que me encontré. No obstante, tomar la decisión no fue fácil, ya que había pasado por dos etapas desgastadoras en el pasado. Al egresar me incorporé al Departamento de Matemáticas, donde colaboré en la formación de sus ingenieros y ayudé en la agotadora labor de su consolidación académica. Así se construían en esa época los nuevos departamentos donde era necesario mejorar el nivel académico, los programas de estudio y obtener los recursos necesarios para crecer. Las dificultades por falta de medios y la escasa ayuda de los decanos era algo que ya habíamos experimentado en el Departamento de Matemáticas. Ahora nos tocaba vivirlo en la División de Computación.

Anteriormente, en 1970, ya había tenido que participar activamente en la reestructuración completa del Plan Común de la Facultad en el ámbito de las matemáticas. La modernización de los planes de estudio -que en esencia aún se mantienen vigentes- se tradujo en tener que impartir materias nuevas y reemplazar a muchos profesores del Pedagógico, quienes eran los principales docentes del antiguo plan. Ese cambio fue liderado por el destacado y carismático profesor Moisés Mellado. Él fue, además, el principal responsable de la creación de la carrera de Ingeniería Matemática, apoyado por el profesor Domingo Almendras.

LAS PRIMERAS DIFICULTADES

Cuando asumí la administración de la División eran tiempos difíciles. Para materializar los planes de desarrollo de la División teníamos problemas constantes con las autoridades por el incumplimiento de los compromisos que adquirían con la Unidad. Las paralizaciones, respaldadas por nuestros alumnos, ayudaron a conseguir los recursos que solicitábamos. Recuerdo especialmente a Fernando Ortiz, quien como representante estudiantil de la carrera, siempre colaboró activamente en todas las iniciativas conducentes a mejorar la calidad de la enseñanza.

Uno de nuestros primeros logros fue el aumento gradual del espacio en planta física. El primer hito fue el traspaso de la Biblioteca del CEC -con todo su personal- desde su ubicación en una casa de Beauchef, al primer piso del edificio de Computación que, poco a poco, pasó a formar parte de nuestra unidad. Con el transcurso del tiempo, la Biblioteca extendió su horario de atención a los días sábado y domingo, usando siempre a los alumnos para la atención en horario no hábil. Luego, la biblioteca comenzó a operar con la modalidad de "estanterías abiertas", es decir, los alumnos podían ingresar al sector de los libros. Hasta entonces, ese era un privilegio del que sólo gozaban los académicos.

EL PRIMER INGENIERO CIVIL EN COMPUTACIÓN

Hasta esa fecha, el DCC tenía docencia especializada en dos programas académicos: el de Magíster en Ciencias mención Computación, creado en 1975, y el de Ingeniería de Ejecución en Procesamiento de la Información, creado en 1969.

Sin embargo, durante años el principal anhelo fue crear un nuevo programa: Ingeniería Civil en Computación. En 1984 nuestra aspiración se haría realidad. La Escuela de Ingeniería había evaluado nuestra propuesta y la decisión final la tomaría el Consejo de Docencia, compuesto por los coordinadores docentes de los departamentos de la Facultad. Nuestro coordinador docente era Pepe Pino y a él le correspondió defender la propuesta. El Subdirector de la Escuela de ese entonces, Isaac Ergas, manifestó que los requisitos se cumplían y, por lo tanto, estaba en manos del Consejo decidir si la Facultad debía o no ofrecer esta nueva carrera.

Sorprendentemente, el coordinador de un departamento, de quien se creía era afín a nuestra propuesta, manifestó rechazo al proyecto. Su principal argumento fue que la computación era una excelente herramienta, pero nada más que eso. En cambio, las ingenierías civiles de la Facultad eran temáticas, orientadas al problema, y convenía mantenerlas así.

Pepe rápidamente contraargumentó que la computación también podía ser considerada un área, y no reconocer su crecimiento en el mundo y en Chile era cegarse a la realidad. Se desató entonces un encendido debate. El coordinador de la carrera de Ingeniería Matemática también vio en el argumento "temático" un ataque a su propia carrera y así surgió un aliado espontáneo.

Finalmente, Isaac Ergas cerró el tema, resumió los argumentos, y dio su propia opinión: la Ingeniería Civil en Computación era una oportunidad para que la Universidad de Chile contribuyera al desarrollo del país en un área importante, de la misma manera como lo había hecho hasta ese momento con las otras especialidades. En la votación, la Ingeniería Civil en Computación fue aprobada por mayoría.

Dos años después, se tituló el primer Ingeniero Civil en Computación. Ronald Corovic aprobó su memoria titulada "Interfaz intuitiva para docencia" en 1986, guiada por el mismo Pepe Pino.

EQUIPAMIENTO UNIX

Teníamos la sensación que nos faltaba algo muy importante en el Departamento. En ese entonces, la vanguardia en computación era sinónimo de Unix. Nuestro esfuerzo se centró en adquirir una de esas máquinas.

Como los VAX eran equivalentes a Unix, realizamos gestiones con Sonda -que distribuía los equipos Digital- para conseguir la donación de un equipo con el sistema operativo Unix. Sin embargo Sonda nunca manifestó interés en esta propuesta, ya que su principal negocio se centraba en el desarrollo de aplicaciones bajo el sistema operativo VMS, de Digital. En esos años, Bell ya había comenzado a cobrar por

> En ese momento no dimensionamos el enorme valor que tenía el haber enviado el primer email del país, ya que esa era una práctica habitual en el resto del mundo. Ahora veo que ese día contribuimos a la historia local.

los binarios de Unix, lo que constituía una traba adicional.

La solución llegó de un modo inesperado. En diciembre de 1982, NCR Corporation sacó al mercado un equipo mediano basado integralmente en Unix: el Tower 1632, que tenía un procesador Motorola 68000, con 512K de RAM. Lo que NCR Corporation hacía, como parte de sus políticas comerciales, era realizar donaciones de estos equipos a universidades y fue así como tuvimos la suerte de recibir uno a fines de 1983.

En una visita realizada en 1984, Gastón Gonnet, destacado profesor del DCC de la Universidad de Waterloo, logró compilar exitosamente Maple en nuestro Tower. Fue así como nuestra Unidad dispuso de un sistema de cálculo simbólico de gran calidad para apoyar la actividad en investigación y docencia.

EL PRIMER EMAIL DE CHILE

En 1985, el Departamento jugó un rol clave en la interconexión vía UUCP entre los equipos Unix de tres universidades: Universidad de Chile, Universidad de Santiago y Pontificia Universidad Católica. Como parte de esta actividad, fue posible el envío del primer correo en Chile entre el DCC, donde estaban Jo Piquer y Patricio Poblete, y la Universidad de Santiago (USACH), donde se encontraban Edgardo Krell y Sergio Mujica.

Recuerdo que Jo Piquer tuvo dificultades no menores para ingresar a la USACH con el módem que haría posible la interconexión. Su aspecto "lana" no cuadraba con el ambiente militarizado que existía en la USACH, que tenía como Rector Delegado a un Coronel de Ejército.

Con UUCP -que funcionaba en una modalidad *store and forward*- cuando se establecía la conexión dial-up se despachaba el correo saliente y se recibía el entrante, esto se repetía en la siguiente conexión. Así, el envío de archivos o correos no era inmediato como lo es hoy con Internet. En ese momento no dimensionamos el enorme valor que tenía el haber enviado el primer email del país, ya que esa era una práctica habitual en el resto del mundo. Ahora veo que ese día contribuimos a la historia local.

Ese mismo año también adquirimos un Tower XP, adquisición que, a pesar de haber sido apoyada por el decanato, tuvo muchas consecuencias molestas, en especial durante el funesto período del Decano interventor de la Facultad, Juan Antonio Poblete.

LA LLEGADA DEL DOMINIO .CL AL PAÍS

Con equipamiento Unix en nuestras instalaciones, y a raíz de una visita que hice en 1986 al centro de investigación INRIA Roquencourt, Francia, solicité ayuda para incorporar al DCC a la red pública internacional UUCP. Como era frecuente en esa época, se nos brindó rápidamente ayuda técnica, y gracias a personas como Yves Devillers, el gurú de UUCP del INRIA, así como del recordado y estimado Philippe Flajolet, fue posible incorporarnos a esa red, vía X.25. Destaco que incluso lo hicimos antes que muchas universidades francesas. Para activar la red UUCP, nuestra contraparte en Chile estuvo conformada por Jo Piquer y Patricio Poblete. En esa época, realizar llamadas de larga distancia y el enganche entre los módems era complejo, por decir lo menos.

Sin embargo, al poco tiempo de ingresar a esta red pública surgió un obstáculo insalvable, Pier Beeterma, responsable en Holanda (CWI) del hub UUCP a nivel europeo, se quejó porque Chile se había incorporado a la red a través de Europa y no vía Estados Unidos como nos correspondía geográficamente. Naturalmente, en el INRIA exclamaron al unísono: *"Ils sont fous ces hollandais!"*[2] como en una historieta de Asterix.

A raíz de esta exigencia, tuvimos que hacer el cambio y nos pusieron en contacto con Rick Adams, fundador de UUNET en Estados Unidos, quién gentilmente colaboró con nosotros para realizar la transición que dejaría tranquilos a los holandeses. Rick, en la famosa máquina Seismo, controlaba la red UUCP en Estados Unidos y también la distribución de News (USENET).

Gracias a la red UUCP, que constituyó un aporte de proporciones para el Departamento, comenzamos a ofrecer, gratuitamente, el servicio de correo a otras universidades chilenas y a algunos organismos públicos y privados. Esta red inició una escuela de administradores de sistemas de gran prestigio. Liderada por Jo Piquer, jugaron un rol destacado Marcelo San Martín, Luis Fuentes y, posteriormente, Eduardo Mercader y Willy Contreras.

En tanto, la distribución de News en Chile, que por su volumen y costos no podíamos recibir en línea, eran despachadas desde Estados Unidos vía cintas. Estos despachos eran esperados con cierta ansiedad ya que Usenet era una verdadera mina de oro de información.

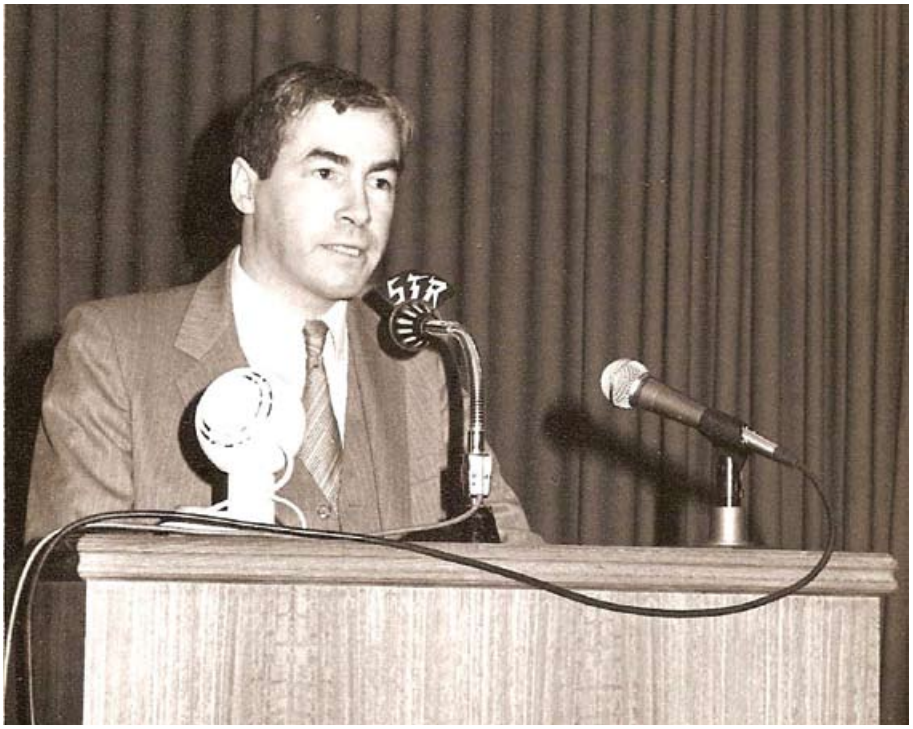
2 "¡Están locos estos holandeses!"

Era una época donde las comunicaciones constituían una pieza clave y las velocidades de transferencia nos significaban verdaderos dolores de cabeza presupuestarios. Teníamos que hacer malabares para pagar las facturas (X.25 al comienzo, telefonía o dial-up después). Recuerdo nuestra alegría cuando llegaron los poderosos módems TrailBlazer de Telebit que nos permitieron comunicarnos a velocidades por encima de los 10.000 bps, algo genial en dichos días.

En un comienzo era habitual reunirse en torno al Tower cuando se realizaban las llamadas, para estar presente cuando se recibía el correo, ya que realizábamos muy pocas llamadas diarias y sólo cuando había algo muy urgente se alteraba el Crontab. En definitiva, ¡era parecido a esperar la diligencia en el viejo Oeste!

Uno de los peligros para las escuálidas arcas del DCC era la recepción de archivos demasiado grandes ya que la lentitud de los módems y el protocolo de comunicaciones hacían necesaria la retransmisión de todo el archivo cuando se presentaba un inconveniente en la transmisión que, como bien se imaginan, ocurría frecuentemente. En una oportunidad, Ricardo Baeza Yates hizo el envío de software para Macintosh, a solicitud de Ricardo Cisternas. Ricardo Baeza Yates, que en ese entonces estaba realizando estudios de Doctorado en Canadá, no pudo imaginar lo que ocurriría con la recepción de su envío. Después de numerosos intentos infructuosos fue necesario recurrir al administrador de UUNET para que eliminara de la cola de envío los archivos que no lográbamos recibir.

Una consecuencia positiva de estar en UUCP fue la llegada, en 1987, de los Top Level Domains (TLDs), ya que desde ese momento pasamos a controlar el dominio .cl que es el que hoy nos distingue. Lo habitual en esa época era que el organismo responsable de UUCP en el país fuese el responsable de la administración de su correspondiente dominio. Ocupé el puesto de coordinador administrativo del dominio y Jo Piquer el de técnico, cargo que aún mantiene en la actualidad. La administración de dominios bajo .cl quedó entonces en manos del Departamento y continúa siendo así.



Jorge Olivos durante una presentación de la empresa Epson (gentileza: Gastón Carreño).

Chile fue el primer país en Latinoamérica en ingresar a UUCP, seguido por poco de Argentina. La incorporación argentina a UUCP ilustra bien cómo ocurrían las cosas en esa época. Dante Caputo, flamante ministro de Relaciones Exteriores del presidente Alfonsín, solicitó al notable y recordado Alberto Mendelzon, en ese entonces profesor de la Universidad de Toronto, Canadá, que informatizara el servicio. Interesaba la interconexión del Ministerio con sus representaciones en todo el mundo, en especial con Estados Unidos. Mendelzon dejó así en funcionamiento una solución basada en UUCP. Al igual que con nosotros, al incorporarse los TLDs, el ministerio argentino heredó la responsabilidad del dominio .ar, algo que a ellos les ha significado más de un problema, ya que en la actualidad todavía continúan entregando el servicio de dns primario para todo el país, esto es, son el NIC argentino. Alberto Mendelzon posteó en las News (USENET), en ese entonces, que Argentina era el primer país de Latinoamérica en ingresar a la red UUCP, posteo que fue rápidamente rectificado por Patricio Poblete, que por esos días se encontraba como profesor visitante en la Universidad de Waterloo.

EQUIPÁNDONOS

En la Facultad contábamos con equipamiento para los alumnos que, en general, podíamos considerar de buena calidad para la época. Hasta hacía muy poco computación era sinónimo de Mainframe y nosotros no éramos la excepción. No puedo olvidar mi desencanto con la plataforma computacional que existía en Francia en el último lustro de los años setenta, ya que en la Facultad disponíamos de un mejor servicio computacional que ellos, basado en Mainframes IBM. A raíz de la política nuclear francesa había restricciones de compra de equipamiento norteamericano y, por lo mismo, un énfasis en el desarrollo de una industria informática propia (Bull). Recién en los ochenta, las universidades francesas fueron autorizadas a adquirir equipos americanos, por ejemplo, VAX y con Unix por cierto.

Nuestros alumnos disponían entre los años 1983 a 1986, para sus trabajos de docencia, de computadores personales con MS-DOS y de terminales con acceso a los equipos Tower, cuando estos entraron en funcionamiento. Por cierto también había facilidades de acceso a los Mainframe IBM.

En 1986 logramos que se compraran equipos Macintosh para apoyar la docencia de Plan Común. Nuestra apuesta fue que en el futuro los computadores tendrían interfaces gráficas, ventanas y harían uso del mouse y así debíamos preparar a los futuros ingenieros. Estos equipos disponían además de un lenguaje Pascal, enseñado en primer año, que facilitaba mucho el debugging de los programas (la manito aquella), algo siempre crucial cuando se enseña un lenguaje de programación.

Al año siguiente, en 1987, logramos adquirir para nuestra docencia de especialidad, los espectaculares computadores personales Amiga 500 de Commodore, un equipo que era multimedial y que acababa de aparecer en el mercado. Tenían componentes muy avanzados para la época, un sistema operativo verdaderamente multitasking, una notable interfaz gráfica, un sistema estéreo de sonido y mouse. Su principal inconveniente era la ausencia de un disco duro. Los desarrolladores de este notable equipo eran todos partidarios de Unix y C. Este equipo fue utilizado durante un buen tiempo, en particular, en las clases de computación gráfica. Recuerdo que el trabajo de memoria de ingeniero de Luis Mateu estuvo basado en una aplicación para Amiga, que tenía la especial particularidad que no se caía nunca y que sirvió por un buen tiempo como herramienta pedagógica en el DCC.

FORMACIÓN DE ACADÉMICOS

Durante el tiempo en que estuve como Director del DCC tuve la oportunidad de hacer uso de los contactos adquiridos en el Departamento de Matemáticas con el área de cooperación técnica de la Embajada de Francia. Recibimos como profesores visitantes a varios investigadores del INRIA entre los cuales destacaría a Philippe Flajolet, Jean Marc Steyeart, Jean-Jacques Lévy, Bruno Salvy, Matthieu Devin y Catherine Granger. La relación fructífera que se creó con el INRIA y que se inició cuando Jean Vuillemin, director de mi tesis de Doctorado, era responsable de un grupo de investigación en esa institución, se mantiene hasta la fecha.

No era posible, en un comienzo, enviar rápidamente a Francia a académicos jóvenes, ya que no dominaban el francés. El primero en salir fue Ricardo Baeza Yates que partió a doctorarse a Waterloo, en Canadá, y lo siguieron otros a Francia como Jo Piquer (con enseñanza media en la Alianza Francesa) y Luis Mateu. Nancy Hitschfeld partió al ETH de Suiza y Nelson Baloian a Alemania.

Aquí se hizo uso de una política de la Facultad que incentivaba la formación de posgrado de sus investigadores jóvenes y que en realidad constituía un prerrequisito para ascender en la carrera académica.

LAS BASES ACADÉMICAS

Conseguir cargos para contratar personal académico fue una tarea compleja. La primera contratación fue Ricardo Baeza Yates y siguieron varias más, entre ellas: Jo Piquer, Luis Mateu, Mario Jofré, Nancy Hitschfeld y Nelson Baloian, entre tantos otros que se me puedan quedar en el tintero, pero que contribuyeron desde su ámbito al desarrollo del Departamento. Una de nuestras preocupaciones permanentes era la dificultad para alcanzar estabilidad con nuestro grupo de investigadores. Así, por ejemplo, ocurrió con Luis Hermosilla y Ernesto Azorín, que después de períodos relativamente cortos como investigadores migraban al sector privado o bien al extranjero en busca de nuevos horizontes.

A mediados de 1986 sufrimos una pérdida muy grande cuando varios investigadores dejaron la Unidad; en dicha oportunidad se fueron Alfredo Piquer, Pablo Alliende, José Benguria, Mario Jofré, Fernando Taboada y, poco después, Rafael Hernández. Si bien Alfredo ya estaba ocupando otros puestos en la Facultad, primero como Director del CEC y después como Director Económico, no esperábamos que fuese a adoptar la decisión de partir a crear una empresa. Esta dolorosa pérdida fue mitigada parcialmente por el hecho de que todos ellos continuaron colaborando en una modalidad part time.

Recuerdo que la última persona que traje al Departamento, poco después de finalizar mi período, fue Gonzalo Navarro, alumno brillante de un curso que impartí en Buenos Aires. Le propuse venir a trabajar a Chile como académico y después de un corto período en la empresa privada en Argentina, aceptó el ofrecimiento.

UN EQUIPO TODO TERRENO

La primera secretaria que contratamos, en 1984, fue Adriana Latorre. En 1985 llegaron Guillermo Morales, como auxiliar, y Margarita Serei como jefe administrativo. Margarita se transformó rápidamente en pieza clave del Departamento por su gran capacidad de trabajo y conocimiento de la operación administrativa de la Facultad. Posteriormente, ingresó Magna Bornand, en 1986, como reemplazo de Adriana que dejó el Departamento para ir a trabajar con Alfredo Piquer, y Sara Quiñones. También contratamos a Fernando Álvarez y Juan Erices como auxiliares. En la Biblioteca ya teníamos a Ximena Rivera, bibliotecaria, y a Fernando Abatte y Gloria Mondaca como personal de apoyo.

Completamos entonces la dotación administrativa del Departamento y disponíamos de todo el primer piso del edificio de Computación.

RECORDANDO A LOS DECANOS

Durante mi período como Director, en la Facultad hubo cuatro decanos. El primero fue Claudio Anguita, poco después asumió Guillermo González, quien no terminó su período y tuvo que sufrir el inicio de las protestas y paros estudiantiles; Juan Antonio Poblete que llegó como Decano Interventor y su período fue de meses (entre abril y octubre de 1985), siendo reemplazado en forma estable por Atilano Lamana, elegido democráticamente por la Facultad.

En general, la relación con los decanos fue de arduo trabajo ya que siempre teníamos dificultades para conseguir los recursos prometidos en el plan estratégico, por el Decano Anguita. La situación no fue más grave gracias a la presencia en el decanato de académicos que comenzaban a jugar el rol de generación de recambio en los puestos directivos de la Facultad. Recibimos un especial apoyo en mi período, primero de Víctor Pérez y después de Francisco Brieva, que en general apoyaban, en la medida de sus posibilidades, nuestra exigencia de un mejor trato por parte de la Facultad.

Cuando sale el Decano González y llega Juan Antonio Poblete, poco después del terremoto de 1985, todo empeoró. Dado el status que teníamos como División y que reseñó Pepe Pino en su anterior artículo, para el nuevo Decano éramos el epítome de las ilegalidades que se cometían en la Facultad, donde en particular existían varios directores "truchos" que inclusive asistían al Consejo de Facultad, entre los cuales me encontraba yo. Con el nuevo Decano llegaron algunas personas que intentaron dividirnos, sin éxito. En este corto período del decanato de Poblete la Facultad inició un movimiento de oposición liderado por Igor Saavedra, profesor del Departamento de Física muy querido y respetado en nuestra Facultad, y como nosotros no dejábamos de lado nuestra propia agenda de reivindicaciones en ocasiones entrábamos en conflicto con la fuerza opositora principal que nos pedía postergar nuestros reclamos.

LAS ÚLTIMAS DIFICULTADES

En ese período fui sumariado y se revisó completamente todo nuestro funcionamiento. Entre los cargos que me imputaron, estaba la solicitud hecha a NCR de hacer llegar el Tower XP, a comienzos de 1985, máquina que ya había sido acordada con el anterior Decano González. Todo el DCC desfiló por las oficinas de la dama responsable del sumario, Directora Jurídica de la Facultad de Medicina. Inclusive llegó a esas oficinas el Director de Matemáticas de ese entonces, Rafael Correa, ya que la División era un apéndice de ese Departamento. Finalmente, me condenaron a una suspensión de un 50% de mi sueldo y el sumario lo dejaron abierto, cerrándose recién cuando regresó la democracia al país, a comienzos de los años noventa.

Cuando asumió Atilano Lamana se restituyó la antigua estructura departamental y regresamos a jugar el rol de Departamento, al menos al interior de la Facultad.

LA HORA DE LA DESPEDIDA

En condiciones anímicas similares a las de Pepe Pino, después de esta tercera tarea realizada en la Facultad, ya estaba en 1988 francamente agotado. Patricio Poblete asumió como nuevo Director.

Después de este período, el DCC disponía de una planta de académicos de tamaño razonable, con formación de Doctorado o en vías de serlo, de una carrera de ingeniería; había resuelto los problemas más acuciantes de infraestructura física; se disponía de un status de Departamento no de jure pero si de facto, y había logrado afianzarse como Unidad en la Facultad. En el ámbito de la investigación se generaban publicaciones y existía una relación privilegiada con el INRIA. Los peligros iniciales de disolución del Departamento habían quedado en el pasado.

Después de este período, el DCC disponía de una planta de académicos de tamaño razonable, con formación de Doctorado o en vías de serlo; de una carrera de ingeniería; había resuelto los problemas más acuciantes de infraestructura física; se disponía de un status de Departamento no de jure pero si de facto, y había logrado afianzarse como Unidad en la Facultad.

Si bien fueron tiempos difíciles, siento que en ese período el equipo de personas del DCC colaboró en sentar las bases de lo que hoy es el nuevo Departamento.

La responsabilidad principal de continuar en la senda de crecimiento quedaba en buenas manos: Patricio Poblete quien siempre actuó como subdirector durante mi período y siempre luchó por recuperar el sitial que correspondía a la disciplina en la Facultad.

AGRADECIMIENTOS

Los recuerdos de este período han sido posibles gracias a la ayuda de Pepe Pino, Patricio Poblete, Jo Piquer y Margarita Serei. BITS

32 años de computación: **de estudiante a Fellow**

"Con Donald Knuth, una de las personas que más admiro, en la ceremonia de premios de ACM, en San Francisco, junio de 2010".

Ricardo Baeza Yates
Vicepresidente de investigación de Yahoo! para Europa, Medio Oriente y Latinoamérica. Profesor Titular DCC, Universidad de Chile (en leave of absence). Catedrático jornada parcial de la Universitat Pompeu Fabra en Barcelona. ACM Fellow, IEEE Fellow y primer socio distinguido de la SCCC.
rbaeza@acm.org
www.baeza.cl

Es difícil escribir sobre la experiencia personal en cualquier tema, especialmente cuando nunca antes se ha hecho y además porque soy, aunque a veces no lo parezca, de naturaleza introvertida. Me considero afortunado de poder trabajar gran parte de mi tiempo en actividades que me gustan, intentando combinar la teoría con la práctica. Para lograr esto, mi principal motivación personal siempre ha sido encontrar mis límites, autogenerando desafíos que han guiado mis pasos.

Por otra parte, es una oportunidad para agradecer a las personas que han influenciado mi carrera y dejar por escrito tanto los recuerdos más relevantes como las lecciones aprendidas, de acuerdo con un criterio muy personal. Como la memoria es frágil, estoy seguro que me he olvidado de eventos y personas tan o más importantes que las que incluyo en este recuento personal. Por lo tanto, pido disculpas de antemano por posibles omisiones. Además, para acotar la extensión de estas líneas, circunscribo mi historia a mi relación con la computación en Chile.

ESTUDIANTE

Cuando en 1979 entré a la Universidad de Chile a estudiar el Plan Común de Ingeniería en la Facultad de Ciencias Físicas y Matemáticas (FCFM) nunca había visto un computador. Tampoco tenía muy claro qué especialidad quería seguir ya que realmente me gustaban la geografía y la astronomía. Mi

primer encuentro con la computación fue ese año e inmediatamente me cautivaron los algoritmos y su lógica implacable. Sin embargo el computador todavía era un ente abstracto pues usábamos las famosas "pantallas de papel" donde escribíamos un programa que luego era ingresado y procesado por un operador, obteniendo el resultado unos días después. Sin embargo, igual decidí seguir ingeniería eléctrica porque era la especialidad más difícil, un primer desafío, y porque aún no existía la ingeniería en computación de seis años. De todas formas tomé el curso de Estructuras de Datos del Bachiller en Computación y luego no pude dejar de tomar las restantes materias, terminando ambas carreras a la vez. Esto generó un segundo desafío que terminó en 1982 cuando reprobé una de las trece materias que había decidido cursar[1].

Finalmente pude conocer directamente un computador alrededor de 1981: un IBM 370. Luego en 1983 usé los primeros microcomputadores que usaban CPM y finalmente, si mi memoria no me engaña, en 1984 llegaron los NCR Tower que usaban Unix, el sistema operativo que uso hasta hoy. En esos años comencé a usar el correo electrónico y procesadores de texto de calidad como *troff*. Durante mis estudios, las tres personas del Departamento de Ciencias de la Computación (DCC) que dejaron una huella importante en mi formación fueron Patricio Poblete, posiblemente el mejor profesor que he tenido, Alfredo Piquer y Jorge Olivos. También tuve la suerte de tener muchos compañeros de generación que luego decidieron ser investigadores, destacando entre ellos a Nancy Hitschfeld y Jo Piquer.

Patricio Poblete fue también mi supervisor de tesis del Magíster en Computación, programa del cual fui el segundo estudiante graduado en enero de 1985, pese a que el programa existía hace ya bastante tiempo. En retrospectiva, mi tesis sobre análisis de algoritmos podría haber sido medio Doctorado y explicaba por qué pocas personas se habían graduado, ya que la exigencia en Chile era mucho mayor, producto de la inseguridad propia de un posgrado pionero en un país pequeño y lejano. Esto ha ido cambiando gradualmente para llegar a una exigencia más acorde con la realidad de países desarrollados. Sin embargo, el esfuerzo tuvo recompensa ya que el resultado principal de la tesis fue publicado en el congreso de la Sociedad Chilena de Ciencia de la Computación ese mismo año. En él participaba Gastón Gonnet de la Universidad de Waterloo al que le gustó mi trabajo y me invitó a realizar el Doctorado con una beca bajo su dirección, el cual comencé en Canadá en mayo de 1986.

El año 1985 fue además importante por otras razones. Primero, fui contratado como instructor del DCC, lo que significaba tener trabajo asegurado al volver del Doctorado. Durante ese año también trabajé en mi Magíster en Ingeniería Eléctrica, que defendí en abril de 1986. La tesis me la dirigió René Nóbile que lamentablemente falleció muy joven unos años más tarde. El tema era realmente de computación, pues abordaba cómo implementar primitivas de computación gráfica en una pantalla orientada a texto, algo que hoy no tiene mucho sentido. Por otro lado esto me impulsó a enseñar durante 1985 uno de los primeros cursos de computación gráfica en Chile.

EL REGRESO

En septiembre de 1989, después de tres años de Doctorado y seis meses de posdoctorado, volví a Chile a los 28 años, promovido a Profesor Asociado gracias a mi productividad científica durante ese tiempo. Es decir, me había saltado la etapa de Profesor Asistente, lo que ahora estimo no fue una sabia idea de parte del comité de evaluación de la FCFM, ya que me salté una etapa de aprendizaje importante. Por otro lado volver a un grupo ya consolidado por el trabajo que habían hecho mis profesores ya mencionados, además de Juan Álvarez y José A. Pino, entre otros, hacía el retorno más fácil y me permitía por fin aportar al desarrollo del DCC, lo que había sido una motivación importante para volver a Chile.

Pero mi país no me recibió exactamente con los brazos abiertos, pues me dio una fiebre tifoidea que me dejó un mes en cama y me cambió la digestión para el resto de mi vida. Poco después, a finales de diciembre tuve la oportunidad de participar en el primer curso de comunicación para la acción que Fernando Flores daba en Chile. Allí no sólo tuve la oportunidad de conocer su filosofía, sino también conocer a mucha gente que más tarde volvería a encontrar, entre ellos a Claudio Orrego.

En 1990 me reincorporé a las tareas docentes, coordinando y reformando el Magíster en Ciencias mención Computación. También, aunque no estaba entre mis áreas de conocimiento, dicté el primer curso de Programación Orientada a Objetos, porque pensaba que era un tema importante que los estudiantes (¡y yo!) tenían que conocer. El siguiente desafío era encontrar fondos para continuar mi investigación en algoritmos de búsqueda en texto, el tema de mi tesis de Doctorado. Así es como conseguí ese mismo año mi primer proyecto gracias a un programa novedoso de la Fundación Andes para trabajar con la industria, en este caso una joven y pequeña empresa, Ars Innovandi, bajo la dirección de Pablo Palma. Este proyecto culminó con el software SearchCity, para Windows 3.1, que permitía buscar dentro de todos los ficheros de un computador personal. SearchCity obtuvo en 1992 el premio de la revista PC Software al mejor software chileno. Sin embargo la idea era demasiado precoz para su época, más viniendo de Chile. Diez años pasarían hasta que una aplicación similar tuviera éxito.

En 1991 obtuve mi segundo proyecto, el primero de Fondecyt, en un tema distinto, visualización de software. En ese momento aprendí la importancia de tener más de una línea de investigación activa y significó el comienzo de una secuencia permanente de proyectos. Ese año también me pidieron ser el organizador de la Conferencia de

1 Física Cuántica, del Bachiller en Física, principalmente por no asistir a clases pues eran muy temprano (si fuera supersticioso habría reprobado por el trece). Por otro lado, los 120 créditos aprobados pasaron a ser parte de las leyendas de la Facultad.

Ricardo Baeza Yates (1994).

> La mayor satisfacción de la labor que uno realiza no son las cientos de citas a tus trabajos científicos, sino el saber que parte de la tecnología que uno ha desarrollado es utilizada por cientos de millones de personas. Para esto es muy importante combinar la mejor teoría con la mejor práctica o, en otras palabras, combinar la investigación básica con la aplicada.

la Sociedad Chilena de Ciencia de la Computación (SCCC) para hacerla realmente internacional. Esto implicaba solicitar por primera vez sólo trabajos en inglés y publicar las actas en una editorial internacional, que fue Plenum Press. Ese congreso lo recuerdo muy bien, no sólo por el nerviosismo de la primera experiencia como organizador, sino también porque nevó en Santiago en octubre y tuvimos que conseguir estufas a última hora, explicándole a los brasileños que habían venido de manga corta, que eso no era la típica primavera santiaguina. Éste fue el primer paso para convertir el Congreso Internacional de la SCCC en uno de los más prestigiosos de Latinoamérica.

En 1992, un año antes de la explosión de la Web, junto con Jo Piquer lideramos la presentación de un proyecto Fondef para crear una plataforma de comercio electrónico. La evaluación fue negativa y frustrante, pues encontraron el proyecto tan bueno, que dijeron que podía ser financiado por fondos privados sin problemas. Fue como correr una maratón, ganarla y en la llegada saber que en realidad estábamos inscritos para los cien metros planos el día anterior. Al parecer Conicyt pensaba que Chile estaba en el Silicon Valley y otra visión futurista se fue al agua.

A finales de 1992 también fui elegido por primera vez Presidente de la SCCC, un reconocimiento tal vez anticipado de mis pares. En este cargo promoví las primeras Jornadas Chilenas de Computación de 1993 en La Serena, que comprendían un foro de investigación tanto nacional como internacional. Esto permitió que desde investigadores consagrados a estudiantes tuvieran un punto de encuentro anual. En estos eventos nacionales conocí a muchos colegas de otras universidades, generando amistades de las que perduran, en particular Leopoldo Bertossi y Miguel Nussbaum.

En 1993 comencé a escribir una columna mensual de divulgación en la *Revista Informática, Crónicas Binarias,* colaboración que duró hasta 2004. También me tocó[2] por primera vez ser Director de Departamento por un período de dos años. No era algo que me gustara hacer pero acepté el desafío con gusto e intente hacerlo lo mejor posible pese a la falta de experiencia. En esta tarea aprendí lo difícil que es mediar entre personas y tuve el apoyo de dos personas que conocían muy bien el funcionamiento del DCC: Magna Bornand y Margarita Serei. Ellas me enseñaron a apreciar los pequeños detalles que hacen funcionar un todo. Agregar a mi currículum una faceta administrativa era un factor necesario para ser finalmente ascendido a Profesor Titular en 1995[3].

En 1994, Jorge Olivos me recomendó traer a un estudiante brillante que había conocido en Argentina y que necesitaba financiamiento para hacer un posgrado. Confiar ciegamente en Jorge fue una decisión acertada pues el estudiante era Gonzalo Navarro. No sólo hizo una tesis de Magíster que ganó el premio CLEI (el segundo de mis estudiantes en conseguirlo), sino que luego fue el primer Doctor graduado en el DCC en 1998, con una tesis excelente que dejó una vara muy alta para mis futuros estudiantes, incluyendo investigadores consagrados como Edgar Chávez en México o Carlos Castillo en España.

En 1995 organicé junto a Eric Goles, el segundo Congreso Latinoamericano de Informática Teórica, LATIN, en Viña del Mar. Éste fue seguramente el primer congreso realizado en Chile publicado por Springer en sus Lecture Notes in Computer Science. También organicé al mismo tiempo el Segundo Workshop Sudamericano en Procesamiento de Palabras, que ahora es un congreso consolidado llamado SPIRE, que se realiza alternadamente en Latinoamérica y Europa. Organicé SPIRE en Chile en 2001 en el Skorpios I navegando hacia la laguna de San Rafael[4] y nuevamente en 2007 en Santiago.

Durante 1995 también comencé a asesorar en nombre del DCC, junto a Patricio Poblete y Jo Piquer, al Servicio de Registro Civil e Identificación para el Proyecto de Nueva Cédula y Pasaporte, donde mi labor estuvo enfocada en la definición y la supervisión de las pruebas del sistema AFIS de identificación de impresiones dactilares. Aquí, por razones obvias, impulsé una versión del pasaporte con hojas extras para viajeros insaciables.

En 1996 tomé mi primer año sabático en Barcelona, que resultó ser un año de renovación y cambios, pues mi transición desde los algoritmos a la aplicación de ellos en la recuperación de información era ya irreversible.

2 Uso este verbo pues como es un puesto que nadie quiere, el turno depende de la jerarquía académica.
3 En ese momento, a los 34 años, era el más joven de la Universidad de Chile. Para los lectores españoles, Profesor Titular en Chile es equivalente a Catedrático.
4 Para muchos de los asistentes, es el mejor congreso en el que han participado, por la forma en que se combinaron el programa científico y el programa social.

CONSOLIDACIÓN

En 1997 me tocó copresidir un congreso de la IEEE Computer Society que era parte de un multievento organizado por Eduardo Vera. Esto me permitió tener los contactos adecuados como presidente del Comité de Programa para publicar las actas del Congreso de la SCCC de ese año en IEEE CS Press, lo que consolidó su posición internacional en la región. Al año siguiente, nuevamente como presidente de la SCCC, incorporé el campeonato de programación de la ACM a las Jornadas Chilenas de Computación. La participación de Chile en este campeonato es un incentivo positivo para los estudiantes de los últimos años y de posgrado.

En 1998 también comencé mi participación en la comisión gubernamental para definir la Agenda Digital durante los Gobiernos de Eduardo Frei y Ricardo Lagos, primero trabajando con Claudio Orrego y luego con Álvaro Díaz, siendo el único representante del mundo universitario. Ese mismo año diseñé y comencé el Postítulo en Gestión Informática. Éste fue el primer programa en Chile que incorporó aspectos humanos que son cruciales en proyectos de software, como liderazgo, trabajo en equipo y negociación.

A mediados de 1999 publiqué en mi sitio Web un manifiesto personal titulado "Diseñemos Todo de Nuevo: Reflexiones sobre la Computación y su Enseñanza", inspirado en ideas de Don Norman y otros, además de mi propia experiencia. Este texto ha tenido influencia en muchos internautas, tanto dentro como fuera de Chile.

En paralelo a estas actividades continuaba con mi trabajo de investigación sobre Búsqueda en la Web, en particular con Nivio Ziviani y Berthier Ribeiro-Neto de la Universidad Federal de Minas Gerais en Belo Horizonte, Brasil. Ellos crearon Akwan, una empresa de tecnología de búsqueda y un buscador para la Web brasileña, TodoBR, en 1999, que fue comprada en 2005 por Google. Con el apoyo de ellos comencé un buscador similar para la Web chilena, TodoCL, en marzo de 2000. Este buscador no sólo permitía experimentar tecnología local y servir a millones de personas al mes (algo que en retrospectiva es mucho más satisfactorio que muchas menos personas citando mi trabajo), sino que me permitió obtener datos que sólo los grandes buscadores tenían, la interacción de cientos de miles de personas y hacer investigación única en el mundo. Este buscador aún funciona y usa tecnología desarrollada en Chile.

Fernando Flores, el Presidente Ricardo Lagos y Francisco Vidal, inaugurando La Ventana Digital (2003).

A finales de ese año organicé el primer Encuentro Sociedad y Tecnologías de la Información que fue todo un éxito y que pude continuar hasta 2009. Este evento permitió difundir distintos aspectos de la computación en la primera década del siglo XXI. Por otra parte, y en paralelo con otros grupos, participé en el desarrollo de la bioinformática en Chile. Primero representé por varios años a nuestro país en la red europea de biología molecular, EMBNET, y luego junto a Juan Asenjo comenzamos en 2001 un curso de bioinformática en la FCFM. Gracias a esto, unos años más tarde lideré el laboratorio de bioinformática de uno de los proyectos del Programa Genoma Chile.

A finales de 2002 me tocó nuevamente ser Director de Departamento. Esta vez tenía la experiencia necesaria pero la situación del DCC era más difícil. La solución fue tener un mecanismo de gobierno parlamentario, con un Consejo de Departamento que tomaba las decisiones y donde el Director era sólo el informador y ejecutor de las mismas. Este cambio, que se mantiene hasta ahora, ha ayudado a fortalecer el desarrollo del Departamento.

En 2002 también lideré la propuesta del Núcleo Milenio Centro de Investigación de la Web, que dirigí hasta 2005 y cuyo financiamiento fue renovado hasta el 2008. Éste es el único núcleo de la Iniciativa Científica Milenio que ha existido hasta la fecha y el proyecto de mayor financiamiento otorgado en Chile en el área de computación. Este proyecto permitió crear una masa crítica de investigadores, financiar estudiantes de posgrado y ser un referente internacional en el área de la Web y la manipulación y búsqueda de información. Actividades realizadas en el marco de este proyecto incluyeron el primer y cuarto congreso latinoamericano de la Web, LA-WEB, los años 2003 y 2007 en Santiago, concursos estudiantiles y una novedosa experiencia de sensibilización social del potencial de Internet para videoconferencias públicas llamada *La Ventana Digital*.

La Ventana Digital consistía en una proyección de video con múltiples canales de audio que comunicaba el Patio de los Naranjos, en el Palacio de la Moneda, con la Plaza de Armas de Arica a dos mil kilómetros de distancia. Esta experiencia fue inaugurada por el Presidente Ricardo Lagos en noviembre de 2003 y tuvo una alta repercusión mediática al permitir a muchas personas conversar gratuitamente por un lapso de diez días con parientes o amigos que no habían visto en años, sin necesidad de entender Internet.

En 2003 también fui honrado con el nombramiento de miembro correspondiente de la Academia Chilena de Ciencias, siendo la primera persona de computación, que trabajaba en Chile, en obtener este rango.

Con el Presidente Ricardo Lagos y Carlos Álvarez (CORFO), junto a Usama Fayyad (Chief Data Officer) y Prabhakar Raghavan (Director de Investigación) de Yahoo!, anunciando el Laboratorio (2006).

Con Víctor Pérez (centro), Rector de la Universidad de Chile; Francisco Brieva, Decano de la FCFM; Ron Brachman (izquierda), VP de Operaciones de Yahoo! Research, y Roberto Alonso (derecha), Director de Yahoo! Latinoamérica, inaugurando el Laboratorio (2006).

Esto ratificaba a la computación como una ciencia en Chile, una antigua y válida aspiración de nuestra comunidad.

A finales de 2005, con un grupo de amigos y discípulos de Alberto Mendelzon, comenzamos a planear una reunión para honrar su recuerdo. Esto se concretó en noviembre de 2006 con el primer Alberto Mendelzon Workshop (AMW), navegando nuevamente por fiordos chilenos en el Skorpios I. Este Workshop es ahora un importante evento regional en Bases de Datos.

Entre 2003 y 2005 tomé un segundo año sabático distribuido entre Stanford, Melbourne, Sydney, Christchurch (Nueva Zelanda) y, nuevamente, Barcelona. En esta última ciudad, Yahoo! me ofreció montar dos laboratorios de investigación, uno en Barcelona y otro en Santiago, un desafío imposible de rechazar. El laboratorio de Santiago se anunció en enero de 2006 y luego se inauguró formalmente en noviembre de ese año. Sin duda el Centro de Investigación de la Web y la experiencia con TodoCL, fueron dos de los factores principales para que Yahoo! instalara un laboratorio de investigación en Santiago, en el cual he contado con el apoyo fundamental de Mauricio Marín. El laboratorio ha sido usado varias veces por CORFO como ejemplo para fomentar otras inversiones extranjeras en tecnología.

Por esta razón en los últimos años no he podido estar de forma permanente en Chile, pero siempre intento estar presente y por eso paso más de un mes al año en el país. Pese a este alejamiento físico parcial, he seguido participando en distintos ámbitos gracias a Internet y he seguido recibiendo reconocimientos en Chile, lo que agradezco profundamente.

EPÍLOGO

La mayor satisfacción de la labor que uno realiza no son las cientos de citas a tus trabajos científicos, sino el saber que parte de la tecnología que uno ha desarrollado es utilizada por cientos de millones de personas. Para esto es muy importante combinar la mejor teoría con la mejor práctica o, en otras palabras, combinar la investigación básica con la aplicada. Esta es una opinión muy personal, pues muchos científicos sólo hacen investigación básica y no creen importante resolver un problema real. Por supuesto la investigación básica es necesaria, pero no es muy productiva si no existen investigadores que permitan aplicar estas ideas para generar tecnología que pueda ser usada por la sociedad en general. Otra satisfacción importante es el impacto indirecto de las personas que uno ayuda a formar, tanto en el ámbito científico como en el profesional.

¿Qué he aprendido? Primero que para desarrollar tu potencial tienes que encontrar tus límites. Segundo, que es mejor no hacer planes pues eso es limitarte a ti mismo[5]. Mejor ir aprovechando las oportunidades que vas encontrando en el camino y que aparecen gracias a tu esfuerzo, pues al final la vida es un *algoritmo online*. Por este último hecho, no te arrepientas de ninguna de tus decisiones, pues seguro que las pensaste bien y siempre más tarde tendrás información que antes no tenías. Tercero, cuando no sepas cómo seguir, reinvéntate todas las veces que sea necesario. En investigación esto significa inventar tú mismo el problema o, por qué no, cambiar de tema. Cuarto, conoce gente, colabora y trabaja grupalmente sin esperar nada a cambio, así se construye tu reputación. Finalmente lo más importante: no pienses en lo que tienes que hacer, ¡sólo hazlo! BITS

5 Personalmente, si hubiera intentado ponerme metas, nunca habría imaginado todo lo que he hecho y el reconocimiento que he obtenido.

COMPUTACIÓN Y SOCIEDAD

El primer computador digital en Chile: **Aduana de Valparaíso, diciembre de 1961**

Plaza Sotomayor, Valparaíso, 1960. **Foto: Mario Ortega P.**

Juan Álvarez Rubio
Académico DCC, Universidad de Chile. Master of Mathematics (Computer Science), University of Waterloo. Ingeniero de Ejecución en Procesamiento de la Información, Universidad de Chile.
jalvarez@dcc.uchile.cl

Hace medio siglo en la Aduana de Valparaíso se instaló el primer computador digital en Chile: un IBM-1401 con 4K de memoria, una lectora/perforadora de tarjetas y una impresora. Su instalación fue motivada por un Tratado de Libre Comercio que comprometió a nuestro país a entregar oportunamente las estadísticas de importaciones y exportaciones. Esta primera experiencia computacional continuó la "mecanización" de apoyo a la administración que comenzó en la Aduana a fines de los años veinte.

INTRODUCCIÓN

La Administración de Aduanas fue organizada en Santiago en 1744 por el Gobernador Agustín de Jáuregui. En 1810, tras la promulgación de la libertad de comercio, se dictó la primera Ordenanza de Aduanas y se estableció la revisión de las mercancías en Valparaíso. Posteriormente, en 1831 la Aduana Mayor o Superintendencia de Aduanas se trasladó desde Santiago a Valparaíso, y desde 1936 funciona en el edificio ubicado en la Plaza Sotomayor [1].

Un primer esfuerzo por "mecanizar" las estadísticas de importaciones y exportaciones se registró en 1927 con la importación e instalación de equipos IBM antes que el fabricante se instalara en Santiago en 1929 y abriera una sucursal en Valparaíso en 1931[2].

Posteriormente, en los años cuarenta y cincuenta el servicio de Aduanas incorporó progresivamente máquinas IBM de "registro unitario" (Unit Record), especializadas en realizar distintos procesos con información perforada en tarjetas: perforadoras, verificadoras, clasificadoras, intercaladoras, tabuladoras, interpretadoras, calculadoras y reproductoras.

EL TRATADO DE LIBRE COMERCIO Y LA REORGANIZACIÓN DE LA ADUANA

El 2 de Mayo de 1961 se promulgó como Ley de la República el Tratado de Montevideo, que instituyó la Asociación Latinoamericana de Libre Comercio (ALALC) que tuvo "por objeto contribuir a la aceleración del desarrollo económico equilibrado de América Latina, a su progresiva industrialización y a la tecnificación de su agricultura y demás actividades primarias, con el fin de promover la elevación del nivel de vida de sus pueblos". El tratado estableció un régimen de gravámenes a la importación de mercaderías procedentes de Argentina, Brasil, Colombia, Ecuador, México, Paraguay, Perú y Uruguay [3].

El Tratado gatilló "la necesidad de convertir a las Aduanas en instrumentos ágiles y operantes de regulación de los tráficos a favor de la liberación y crecimiento de los mercados complementarios supranacionales". En consecuencia, en 1961 se procedió a dar "una nueva conformación de la Superintendencia de Aduanas, consultando en su mecanismo los Departamentos, Secciones y Funciones que requiera la Dirección del servicio para una administración eficaz y responsable de las Aduanas, en consonancia con las modernas técnicas y principios que gobiernan la organización racional"(...) "Entre las obras de mayor entidad, cabe destacar la modernización del Departamento de Estadística de la Superintendencia de Aduanas, con el fin de que esté en condiciones de programar y procesar a tiempo las estadísticas del comercio internacional y de cabotaje, como instrumento de información y evaluación económica para las autoridades gubernativas, organismos públicos y demás entidades vinculadas a la planificación o estudio de nuestras relaciones comerciales con terceros países" [4].

La reorganización del Servicio y el nombramiento de Octavio Gutiérrez como nuevo Superintendente mereció el titular principal de portada del diario La Nación [5] e incluso un editorial [6].

Primeros equipos en la Aduana a fines de los años veinte (gentileza: La Nación).

LA NECESIDAD DE UN COMPUTADOR Y LA PREPARACIÓN PREVIA

Las nuevas necesidades estadísticas derivadas de la ALALC y de la implantación de la nomenclatura de Bruselas aconsejaron la instalación en el Departamento de Estadística de un computador IBM-1401 comercializado internacionalmente desde 1959 para aplicaciones administrativas o "comerciales", es decir, para procesos con un gran volumen de datos pero con cálculos sencillos (sumas, promedios, porcentajes, etc.).

La Aduana decidió aceptar la oferta de arriendo de un 1401 y convocó a sus empleados a un concurso interno para capacitarse como programador. Se presentaron alrededor de sesenta empleados que fueron sometidos a un test de aptitudes aplicado por IBM. Finalmente fueron seleccionados René Cabezas, de 26 años, Jefe de Máquinas UR, y Leopoldo Valdivia, de 27 años. Los dos fueron entrenados en diagramas de flujo y en el lenguaje ensamblador SPS (Symbolic Programming System) por Federico Cavada del Departamento de Ingeniería de Sistemas de IBM, ubicado en calle Prat 772 de Valparaíso. Para su labor de soporte de sistemas, Cavada fue previamente capacitado por Hernán Carvallo de IBM, quien a su vez había recibido entrenamiento en México en 1959.

Una vez terminada la capacitación, y bajo la jefatura de Domingo Godoy en el Departamento de Estadística, René Cabezas y Leopoldo Valdivia comenzaron a programar un sistema estadístico de importaciones. Terminados los programas, escritos con "papel y lápiz" en hojas de codificación, en el mes de agosto de 1961 fueron autorizados, con una asignación especial de E°200, para viajar a Buenos Aires para realizar pruebas en las instalaciones de IBM-Argentina, acompañados por Alfonso Carvallo de IBM [7]. Los programas fueron perforados en tarjetas y sólo al final de la primera semana lograron que funcionaran y en las dos semanas siguientes los corrigieron y afinaron.

INSTALACIÓN DEL COMPUTADOR IBM-1401

El diario La Nación del lunes 11 de diciembre de 1961 consignó en una breve noticia la llegada del primer computador digital a Chile:

Computador electrónico instalarán en la Aduana

Prestará sus funciones en el Departamento de Estadística de estos servicios.

En el vapor "Imperial" de la Sudamericana llegó a Valparaíso el nuevo sistema

Computador Electrónico 1401, que será instalado en el Departamento de Estadística de la Superintendencia de Aduanas, según convenio suscrito con la IBM. Este equipo es único en su género y se trata del primero en llegar a Chile.

La instalación del computador Electrónico 1401 ha sido determinada por las nuevas necesidades estadísticas derivadas del funcionamiento de la Asociación Latinoamericana de Libre Comercio, de la implantación de la nomenclatura arancelaria de Bruselas, de un control más afinado de los valores de las mercaderías de importación y otros programas relacionados con los compromisos internacionales contraídos por Chile.

En el edificio de la Superintendencia quedaron ya totalmente terminadas las obras previas a su instalación. Los trabajos fueron ejecutados por el contratista señor Madrid.

Inventarios

Desde 1962, por otra parte, los inventarios de bienes muebles serán llevados en forma automática por medio de fichas perforadas. Esto permitirá mantener un registro centralizado de inventarios en la Dirección del Servicio, que abarcará todas las dependencias del país.

El computador 1401 se instaló en el primer piso del Edificio de Aduanas. Al respecto, el contrato de arriendo establecía que los impuestos de importación debía pagarlos el cliente. En consecuencia, y para evitar los elevados aranceles, el edificio fue declarado como parte del recinto aduanero. La instalación fue registrada en el boletín de la Superintendencia de Aduanas del mes de enero de 1962:

Mecanización del Servicio de Aduanas

Se ha instalado en el Departamento de Estadística de la Superintendencia el moderno computador electrónico 1401, equipo que es el más moderno que existe en el país. Su instalación fue determinada por las necesidades estadísticas originadas por el funcionamiento de la Asociación Latinoamericana de Libre Comercio, la implementación de la Nomenclatura Arancelaria de Bruselas, un control más afinado de los valores de las mercaderías de importación y exportación y otros programas relacionados con los compromisos internacionales suscritos por Chile.

Para examinar en el más alto nivel administrativo y técnico las bondades de su aplicación en la Aduana se efectuó un seminario para ejecutivos del Servicio, curso que fue dictado por el experto de la IBM, señor Alfonso Carvallo Díaz.

Se ha contratado además, otro equipo electrónico para mecanizar las secciones Liquidación y Control de la Aduana de Valparaíso, el cual estará instalado en 1963. Este computador podrá efectuar en sólo tres horas el total del trabajo correspondiente a una jornada diaria, con el consiguiente mejor y más rápido servicio a la industria y al comercio.

Se está examinando, por otra parte, la conveniencia de mecanizar el control de pasajeros a las zonas liberadas sobre la base de tarjetas perforadas. Esto permitirá controlar con prontitud y seguridad los viajes efectuados a esas zonas por cualquier persona.

Confirmando la importancia de instalar el primer computador, viajó a Chile nada menos que el presidente de IBM Arthur K. Watson. El diario La Nación del 15 de enero de 1962 dio cuenta del hecho:

Uno de los mayores productores de equipos electrónicos llegará hoy

Es el presidente de la I.B.M. que opera en 92 países y que ha comenzado a colocar con éxito sus computadores electrónicos en la Administración chilena

"La IBM cuyos computadores electrónicos comienzan a revolucionar el sistema de trabajo de las más importantes oficinas chilenas, liberando al personal de todas las tareas de carácter rutinario y multiplicando el tiempo creador, está de pláceme pues será visitada por el más alto ejecutivo de la Compañía, su presidente, Arthur K. Watson, quien llega hoy, a mediodía, en el


...un corte de género regalado

Diario La Nación del 11 de diciembre de 1961.

Jet de Panagra, procedente de Nueva York", informó a LA NACIÓN el gerente general en nuestro país, Hernán Elizalde.

Numerosas Condecoraciones

"El señor Watson, que fue condecorado en una oportunidad por el gobierno de Chile con las insignias de la Orden al Mérito "Bernardo O'Higgins" por prominentes servicios a nuestro país, es un distinguido hombre de negocio norteamericano, del Estado de Nueva York. Graduado en la Universidad de Yale, sirvió al ejército de su patria durante la última guerra, mereciendo el ascenso a Mayor..."

Franco éxito en Chile

"La I.B.M. que ya opera en 92 países, está prestando en Chile un positiva cooperación a la modernización de nuestros clásicos sistemas de trabajo, particularmente en el campo de la Administración y de los negocios. Más de ochenta clientes de nuestros equipos de máquinas eléctricas y electrónicas de contabilidad, control y estadística, están funcionando ya con pleno éxito.

Se encuentra ya en la Aduana de Valparaíso el primer computador electrónico I.B.M., que ha llegado al país y que va destinado a la Superintendencia del mismo servicio,

en Valparaíso. Impuestos Internos ya cerró contrato por un equipo electrónico similar. En general, el interés del comercio, la industria y la Administración es tan grande por aprovechar las ventajas de la electrónica al servicio de la racionalización de las tareas que creemos que en los próximos dos años pasarán de una docena de computadores que colocaremos en Chile".

La I.B.M. según información de Elizalde tiene 3 sucursales y en ella trabajan 296 empleados en su totalidad nacionales.

La puesta en marcha del 1401 no fue una tarea sencilla y tardó cerca de tres meses. Su inauguración quedó consignada en el diario La Nación del miércoles 14 de marzo de 1962:

Un moderno equipo electrónico inauguró aduana de Valparaíso

Valparaíso.- Uno de los más modernos y grandes computadores electrónicos actualmente en uso en el país, comenzó a prestar servicios en la mañana de ayer en el Departamento de Estadísticas de la Superintendencia de Aduanas.

La máquina fue proporcionada por la IBM, empresa que también dictó los cursos correspondientes a los funcionarios que tendrán a su cargo el manejo de la moderna máquina.

El funcionamiento del computador permitirá a la Aduana agilizar sus labores en el aspecto estadístico, ya que operaciones que anteriormente demoraban varios meses, ahora será posible realizarlas en el término de pocos días.

El equipo inaugurado ayer consta de tres máquinas: la unidad de proceso o "pensante" (cerebro electrónico), la parte encargada de la lectura y perforación de tarjetas, y una impresora de alta velocidad.

CARACTERÍSTICAS TÉCNICAS

El computador IBM-1401 contaba de un procesador con 4K de memoria, una lectora/perforadora 1402 de 400 tarjetas por minuto y una impresora 1403 de 600 líneas por minuto. El computador no disponía de un sistema operativo, por lo que era operado por los propios programadores. De hecho, una compilación entregaba como resultado un programa en lenguaje de máquina perforado en tarjetas. Para su ejecución, las tarjetas debían ser trasladadas manualmente a la lectora, activando su lectura y ejecución a través del panel de control.

Lecto-perforadora 1402, procesador 1401 e impresora 1403 similares a los de la Aduana.

Una vez que el computador estuvo operativo, el sistema estadístico se procesó paralelamente, tanto en las máquinas UR, como en el 1401. Después de jornadas de trabajo que abarcaron varios días completos, se logró tener las estadísticas actualizadas, logrando superar el retraso de dos años del sistema antiguo. Cabe señalar que los resultados se imprimieron en alrededor de 400 páginas de formulario continuo.

La máquina resultó bastante robusta y para el soporte de hardware, IBM destinó a los técnicos Carlos Fuentes y Lautaro Medina. Sólo se recuerdan problemas intermitentes con la lectora de tarjetas que resultó con algunos daños durante el desembarco que se atribuyeron a "la falta de una paloma para el operador de la grúa". Por otra parte, jamás se descubrió la razón de algunas "caídas" inexplicables del procesador, aunque se observó que coincidían con los movimientos de algunas maquinarias del muy cercano recinto portuario.

La IBM, cuyo encargado de la sucursal en Valparaíso era Carlos de la Barrera quien

René Cabezas compilando un programa (gentileza: René Cabezas).

había trabajado anteriormente en la Aduana, utilizaba el computador arrendado a la Aduana para mostrarlo a sus clientes. En una oportunidad uno de los visitantes presionó uno de los botones del panel de control preguntando "¿para qué sirve esta tecla?" Resultado: se interrumpió abruptamente un programa que llevaba horas de proceso. Desde entonces, la Aduana restringió las visitas de demostración a los clientes de IBM.

EVOLUCIÓN POSTERIOR

Considerando la satisfactoria experiencia inicial, que incluso significó recibir una carta de felicitación del presidente Jorge Alessandri, en 1963 la Aduana decidió "agrandar" el computador agregando 4K de memoria y cuatro unidades de cinta magnética 729. Como anécdota se puede señalar que se recibió un programa que ordenaba ("sorteaba") una cinta, pero lamentablemente no funcionó. Sorprendentemente, después de examinar los cientos de tarjetas que contenían el programa en el lenguaje de máquina, René Cabezas logró corregirlo. Al respecto, un "sort" con el algoritmo de cascada tardaba alrededor de doce horas en ordenar la información de una cinta magnética de 2.400 pies de longitud grabada con una densidad de 800 bits por pulgada.

La incorporación del computador tuvo efectos en el Departamento de Estadística. En diciembre de 1962 se estableció una planta de ocho personas para la sección de máquinas de contabilidad y estadística: un jefe de máquinas, tres programadores, tres operadores y un jefe de registro y despacho[8]. Se decidió entonces capacitar más programadores. Después de un nuevo test de habilidades se seleccionó a Guillermo Fliess, Luis Reyes, Luis Prado y Raúl Domínguez. Esta vez acudieron a las oficinas de IBM en Santiago, ubicadas frente a la Estación Mapocho, donde recibieron cursos de capacitación en el lenguaje de máquina y en el lenguaje simbólico Autocoder. Bajo la nueva jefatura del Departamento de Estadísticas de don Carlos Reyes Lanyon desarrollaron aplicaciones estadísticas, un sistema de remuneraciones y una base de datos de personal. Por otra parte, IBM instaló otro computador 1401 para la Armada de Chile, a unas pocas cuadras de distancia, estableciéndose un convenio de respaldo mutuo.

Edificio de Aduanas, agosto de 2010.

En 1965, durante el gobierno de Eduardo Frei Montalva, se realizó una asesoría externa que dirigió el experto norteamericano Robert Kennedy[9]. Como resultado, la Aduana nuevamente se reestructuró y se creó el Centro de Procesamiento de Datos a cargo de Guillermo Fliess y se desarrollaron nuevos sistemas aduaneros. Posteriormente, en 1973 se agregaron al computador 24K de memoria y una unidad de disco. Finalmente, en 1975 el IBM-1401 fue reemplazado por un IBM-370 modelo 125, uno de los primeros computadores de ese tipo en Chile.

CONCLUSIONES

La instalación del primer computador en Chile en la Aduana de Valparaíso en 1961 presenta elementos de continuidad y cambio. La continuidad se reflejó al considerar el computador como un hito más en la "mecanización" de apoyo a la administración que comenzó a fines de los años veinte. La continuidad también

> La Aduana de Valparaíso ocupa un lugar privilegiado en la Historia de la Computación en Chile. En 1961 se programaron las primeras aplicaciones administrativas y se recibió el primer computador digital en Chile.

De izquierda a derecha: Federico Cavada, René Cabezas y Guillermo Fliess.

se reflejó en la mantención, tanto del proveedor de los equipos (IBM), como la estructura administrativa de la institución (sección de máquinas del Departamento de Estadística). Un reflejo de la percepción de continuidad fue la escasa y anecdótica cobertura noticiosa por parte de la prensa escrita.

Los elementos de cambio fueron percibidos con la obtención oportuna de los primeros resultados y la comprobación de la presencia de una tecnología poderosa y flexible. El efecto fue la creación en 1965 de un Centro de Procesamiento de Datos transversal a toda la institución y que fue uno de los elementos fundamentales en la restructuración de la Aduana. En síntesis, la experiencia computacional de la Aduana fue valiosa y aleccionadora para otros servicios e instituciones del Estado y para el desarrollo futuro de la disciplina.

EPÍLOGO

La Aduana de Valparaíso ocupa un lugar privilegiado en la Historia de la Computación en Chile. En 1961 se programaron las primeras aplicaciones administrativas y se recibió el primer computador digital en Chile. Este medio siglo de Historia merece al menos una placa conmemorativa virtual, análoga a las que se encuentran actualmente en la fachada de su edificio y que rinden homenaje a un grupo de sus trabajadores y al poeta nicaragüense Rubén Darío, quien trabajó en la Aduana a fines del siglo XIX, período en que escribió su obra más importante: "Azul" en 1888.

AGRADECIMIENTOS

Aunque la responsabilidad de la redacción es exclusivamente del autor, agradecemos la valiosa colaboración de Guillermo Fliess, René Cabezas y Federico Cavada, quienes proporcionaron valiosa información y fotografías de la época. Gracias también a mi colega en el proyecto "Historia de la Computación en Chile" Claudio Gutiérrez por sus comentarios.

Guillermo Fliess tiene actualmente 73 años y acaba de cumplir 51 años trabajando en la Aduana. René Cabezas, 74 años, trabajó en la Aduana hasta 1963 y permaneció activo en el área hasta el año 2008 y vive actualmente en Santiago. Federico Cavada, 82 años, trabajó en IBM hasta su retiro y vive actualmente en Viña del Mar.

Agradecimientos también para Gabriel Ahumada de la Biblioteca del Congreso; Carlos Adriazola del Archivo del diario La Nación; Patricia Liberona del Archivo Central Andrés Bello; Ana María Carter, Daniel Encalada y Luis Cortés de la Biblioteca de la Facultad de Economía y Negocios de la Universidad de Chile. BITS

REFERENCIAS

[1] Aduana de Chile, "Historia de la Aduana de Chile", septiembre 2011. http://www.aduana.cl/prontus_aduana/site/artic/20070224/pags/20070224173229.html

[2] IBM de Chile, "80 años IBM Chile"; septiembre 2011. http://www-03.ibm.com/marketing/cl/marketing/historia/index.shtml

[3] Superintendencia de Aduanas de Chile, "Boletín Oficial", diciembre 1962.

[4] Superintendencia de Aduanas de Chile, "Boletín Oficial", abril 1962.

[5] La Nación, "Reestructuración total de las Aduanas – Instrucciones impartió Ministro de Hacienda al nuevo jefe del Servicio", portada del 21 de diciembre de 1961.

[6] La Nación, "El Servicio de Aduanas", editorial del 23 de diciembre de 1961.

[7 Superintendencia de Aduanas de Chile, "Boletín Oficial", agosto 1961.

[8] Superintendencia de Aduanas de Chile, "Boletín Oficial", diciembre 1962.

[9] Superintendencia de Aduanas de Chile, "Boletín Oficial", abril 1965.

OPEN DATA

Open Data: nuevo **paradigma en el manejo de datos**

Alejandro Barros
Magíster en Ciencias mención Computación, Universidad de Chile. Director de e.nable. Ex Secretario Ejecutivo Estrategia Digital de Chile 2007-2008. Consultor internacional de empresas e instituciones públicas, especializado en planificación estratégica tecnológica, políticas tecnológicas, gobierno electrónico, compras públicas e introducción de tecnologías en procesos de negocios. Académico Asociado del Centro de Sistemas Públicos de la Universidad de Chile.
abc@alejandrobarros.com
www.alejandrobarros.com

El desarrollo digital de los Estados se inició buscando aumentar la eficiencia del quehacer de estos, digitalizando procesos y actividades con el objetivo de mejorar su desempeño y reducir los costos de operación, el foco eran los procesos de back office.

El desarrollo del gobierno electrónico, ha estado identificado con un modelo de gobierno-céntrico. Desde hace algunos años, su eje ha cambiado, poniendo al ciudadano al centro, lo que denominamos modelo ciudadano-céntrico. Este cambio se sustenta en una mejor identificación de su misión de servicio al ciudadano, producto de la madurez alcanzada y en parte por el aumento del nivel de empoderamiento que los ciudadanos han adquirido en los últimos años.

Lo anterior genera dos consecuencias en términos del diseño de las políticas públicas y desarrollo digital de los Estados:

1) **Nuevo modelo de servicios**, los servicios que presta el Estado a sus ciudadanos, sean estos en la forma de personas naturales o instituciones, deben diseñarse e implementarse teniendo como foco quién es el receptor de los mismos y no quién los produce, para lo cual el Estado debe incorporar al momento de

su diseño atributos esenciales de este nuevo enfoque:

a. One-stop-shop (ventanilla única).
b. Múltiples canales de atención.
c. Altos estándares de usabilidad.
d. Interoperabilidad.
e. Niveles de servicio definidos *ex ante*.

2) **Gobierno abierto,** concepto que agrupa la participación, transparencia y colaboración de los ciudadanos en las políticas públicas, en este ámbito el Open Data juega un rol relevante y se ha transformado en la forma de operacionalizar dicho enfoque.

OPEN DATA

Hoy en día, cada vez más países, han adoptado el modelo de gobierno abierto, el cual se sustenta en un cambio de paradigma frente a los datos que están en poder del Estado, transformándolos en públicos y promoviendo su acceso a uso por parte de los ciudadanos. Los Gobiernos de Estados Unidos[1], Australia[2] y Reino Unido[3] por mencionar algunos han definido en los últimos años modelos de gobierno abierto, en particular de Open Data.

Los Estados, producto de su función recopilan y producen grandes volúmenes de datos de todo tipo[4] (climatológico, económico, social, cultural y muchos otros), los cuales generalmente se encuentran con accesos restringidos, en muchos casos ni siquiera se conoce su existencia, en formatos no estandarizados y con una gestión bastante deficiente por parte del Estado.

El enfoque Open Data busca cambiar esto, disponibilizando esa data a todos los ciudadanos, lo cual en el poco tiempo que lleva este tema ha demostrado que genera alto valor público. Daniel Lathrop y Laurel Rume en su libro Open Government plantea tres conceptos que me parecen fundamentales a la hora de entender el impacto del Open Data.

- *La información pública es una forma de infraestructura, con el mismo nivel de importancia que otras infraestructuras (agua, electricidad, carreteras).*
- *Debemos maximizar el valor público a partir de la data existente en manos del Estado.*
- *La magia de datos abiertos es que habilita la transparencia y la innovación.*

CONFUSIONES HABITUALES

La disponibilización de los datos públicos es un proceso que debe estar normado, el cual si bien es bastante nuevo ya se rige por ciertos estándares internacionales. Hoy en día, existen dos marcos de referencia para ello, las recomendaciones sobre información pública de la OCDE[5] y los ocho principios de datos abiertos, esto es:

1. Completitud, toda la data es pública, no sólo lo que la autoridad estime pertinente (preprocesada).
2. Fuente debe ser primaria (raw data).
3. Oportuna, debe disponibilizarse en forma inmediata o al menos con poca demora.
4. Disponible a todo tipo de usuarios (sin restricción de acceso).
5. Procesable, esto es, datos estructurados que puedan ser procesados por un computador.
6. No discriminatoria, disponible a cualquiera sin necesidad de registro.
7. No propietaria, esto es, que no puede estar en formatos asociados a alguna entidad o bien que requieran de algún tipo de herramienta propietaria para su uso.
8. Licenciamiento libre, no sujeto a ningún tipo de copyright, patente u otro tipo de derecho.

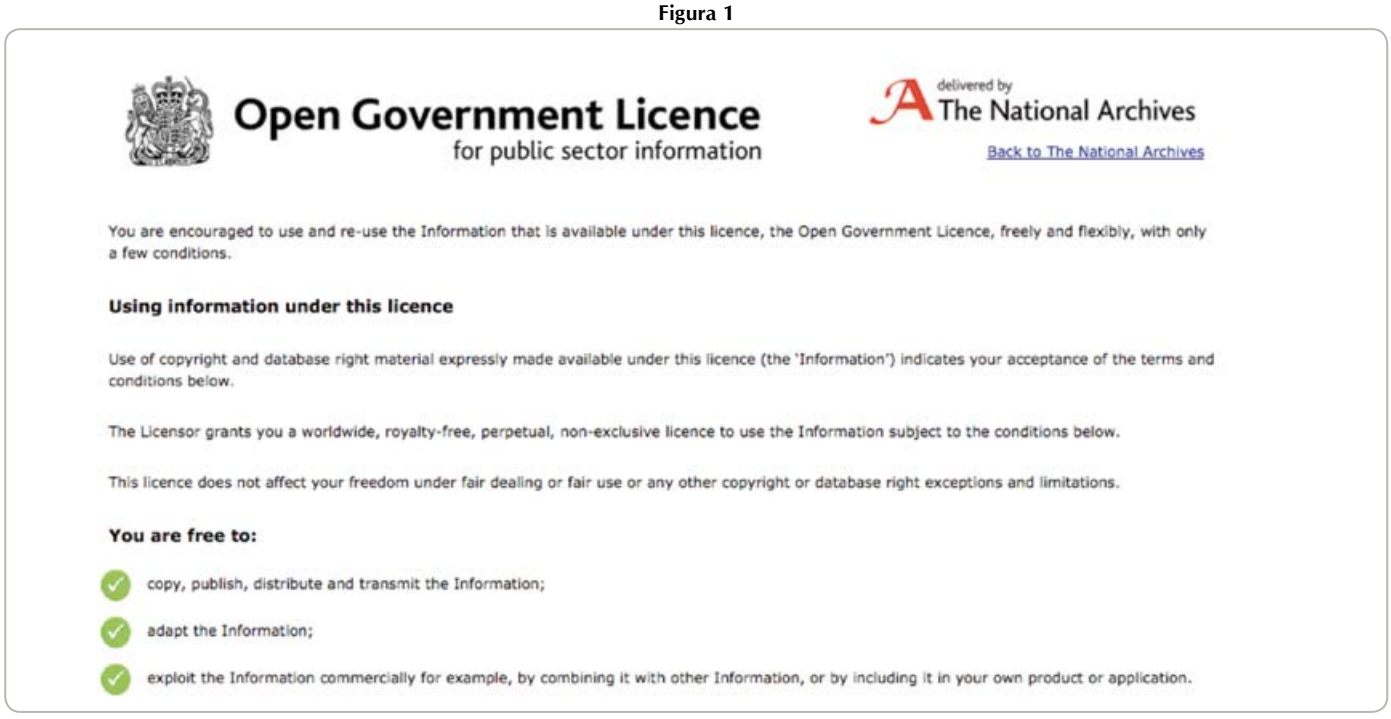
Ahora bien cuando contrastamos estos principios con la realidad de los Estados de la región, es que vemos que existe una gran brecha, el comportamiento de los países de la región está más bien marcado por el secretismo y la opacidad respecto de los datos en poder del Estado.

Un ejemplo reciente de ello fue la discusión que se dio en Chile respecto de los niveles de pobreza utilizando para argumentar los resultados de la encuesta Casen, si el acceso a esos datos hubieran seguido

> La posibilidad de que los ciudadanos puedan acceder a los datos públicos en forma simple va a impactar positivamente en nuestras sociedades, mejorando la democracia y la economía.

1 http://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment/
2 http://agimo.govspace.gov.au/2010/07/16/declaration-of-open-government/
3 http://thenextweb.com/uk/2010/01/20/uk-government-open-data-revolution/
4 The Future of the Goverment, WEF, 2011.
5 http://www.alejandrobarros.com/content/view/718580/Que-hacer-con-la-informacion-del-Sector-Publico-segun-la-OCDE.html#content-top

Figura 1



los principios antes señalados, podríamos haber visto una noticia como la siguiente:

"Un grupo de estudiantes secundarios con la ayuda de su profesor refutaron las conclusiones del ministro de Hacienda respecto de la distribución del ingreso, llegaron a esa conclusión producto del análisis de los datos publicados de la encuesta CASEN 2010".

Diario local, junio 2010.

Como lo mencionaba, una noticia como la anterior no sería fácil, ya que el acceso a los datos de dicha encuesta es bastante tortuoso.

Cuando el Ministerio de Planificación (MIDEPLAN) dice que tiene los datos de la encuesta CASEN, lo primero que un ciudadano se pregunta cuando accede al sitio donde se encuentra dicha información denominado "Casen Interactiva[6]" es:

1) ¿Por qué debo demostrar que soy un investigador como solicita MIDEPLAN para obtenerlo?
2) ¿Por qué tengo que dar a conocer los objetivos de la solicitud de los mismos?
3) ¿Por qué tengo que tener un software específico para procesarlo (SPSS o Stata) o bien bajar un ejecutable con los datos?

Algunos países han avanzado más allá del mero acceso a los datos, el modelo Open Data ya ha demostrado que produce cambios importantes no sólo en el sector público, adicionalmente está generando sinergias en la innovación y el emprendimiento, tal es el caso del Reino Unido, que a través de su organización The National Archives (Archivos Nacionales) ha definido un modelo de licenciamiento de la data[7] el cual permite lo expuesto en la Figura 1.

No sólo como ciudadano puedo acceder a la data pública, sino que además puedo usarla en términos comerciales; este modelo permite que se genere emprendimiento en torno a dichos datos.

CONCLUSIONES FINALES

La posibilidad de que los ciudadanos puedan acceder a los datos públicos en forma simple va a impactar positivamente en nuestras sociedades, mejorando la democracia, producto de mayores niveles de participación y compromiso, mejorando la economía ya que se ha demostrado que el acceso a datos públicos permite realizar obras derivadas a partir de esa data con su consecuente impacto en la innovación y el emprendimiento.

Los Estados de la región ya están tomando cartas en este tema, partiendo por establecer marcos jurídicos, como son las leyes de transparencia que muchos de los Estados de la región han establecido. Estos son tímidos pasos pero van en la dirección correcta.

Algunas iniciativas interesantes en esta materia en nuestro país son las que están liderando la Biblioteca del Congreso y el Consejo para la Transparencia con sus proyectos de Open Data, es de esperar que estas organizaciones contagien al resto del Estado chileno.

Los ciudadanos debiéramos estar alerta y hacer cumplir nuestros derechos en lo que a los datos en poder del Estado respecta, recordándole que en lo que se refiere a la data pública, al Estado le corresponde sólo su administración y que la propiedad de estos es de los ciudadanos en su conjunto, salvo en casos muy excepcionales. BITS

6 http://celade.cepal.org/redatam/paises/chl/mideplanii/Index.html
7 http://www.nationalarchives.gov.uk/doc/open-government-licence/

Open Government Data en el mundo

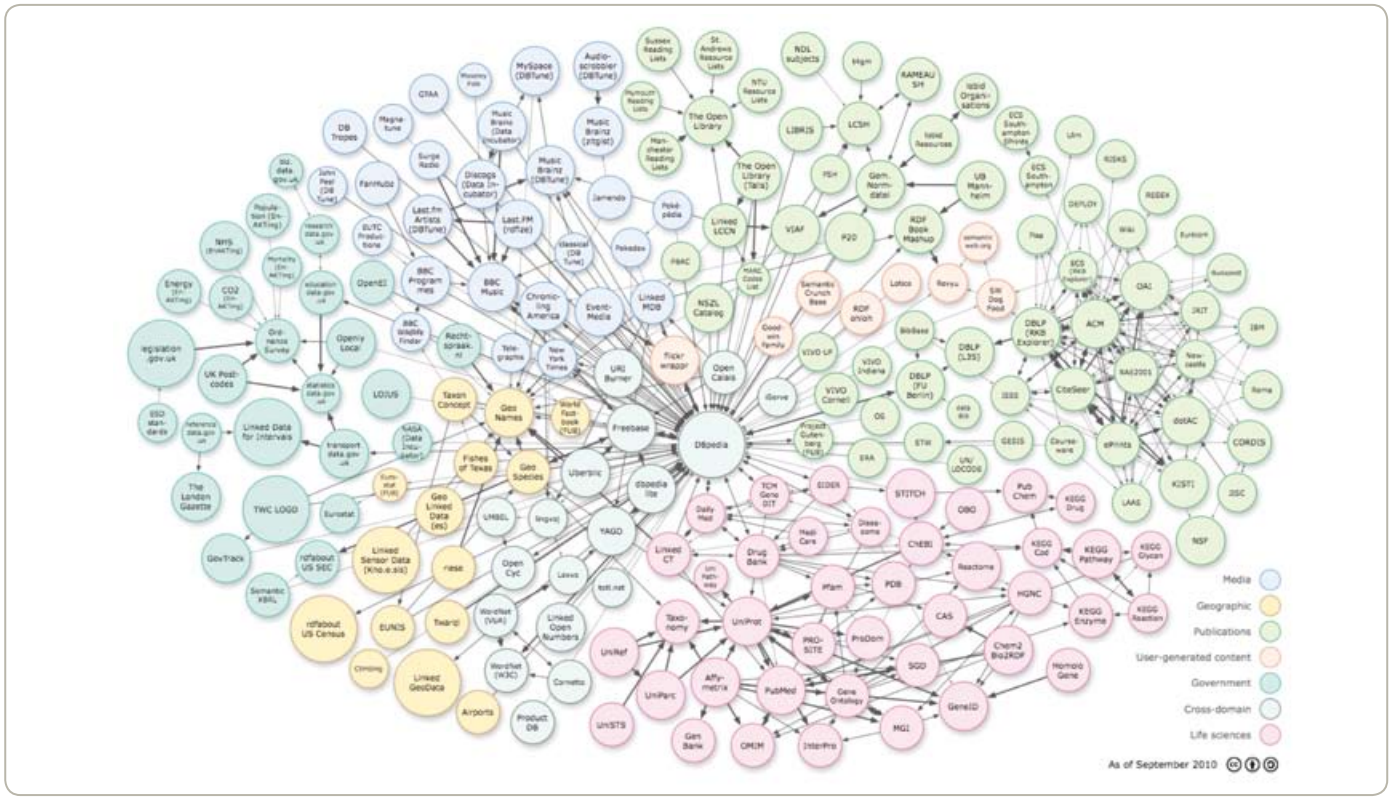
Álvaro Graves
Ingeniero Civil en Computación y Magíster en Ciencias mención Computación, Universidad de Chile. Estudiante de PhD en Cognitive Science, Tetherless World Constellation, Rensselaer Polytechnic Institute, Estados Unidos.
alvaro@graves.cl

Desde hace algunos años ha tomado fuerza la idea de publicar libremente datos de gobierno de distintos países, tanto a nivel nacional, regional y municipal. Este movimiento, conocido como Open Government Data (OGD) se ha extendido durante los últimos años y actualmente más de una veintena de países, incluyendo Estados Unidos y el Reino Unido, implementan portales de publicación de datos. Asimismo, este movimiento se ha visto fuertemente asociado a Linked Data, que consiste en una serie de principios para publicar datos usando tecnologías de la Web Semántica, que los hacen fácilmente procesables por máquinas. Esta simbiosis ha beneficiado a distintas organizaciones al interior del gobierno, así como a académicos, investigadores y ciudadanos en general. El presente artículo describe cómo los países han comenzado a adoptar OGD y cómo el uso de Linked Data ha ayudado en la publicación de datos.

¿QUÉ ES OPEN GOVERNMENT DATA?

Open Government Data consiste en un conjunto de principios que apuntan a que los datos generados o usados por los gobiernos debiesen estar a libre disposición y uso por parte de los ciudadanos. Existen varias razones que justifican esto: en primer lugar, los datos generados por el gobierno son financiados con los impuestos de todos. ¿No deberían todos los ciudadanos poder usarlos, dado que han pagado por ellos? En segundo lugar, el reuso de estos datos permite que otras personas se beneficien directa e indirectamente de estos, aumentando su valor y utilidad. En Estados Unidos, empresas como BrightScope.com (que reporta información sobre consejeros financieros) y aplicaciones como Roadify.com (que entrega información sobre transporte público de Nueva York en tiempo real) utilizan

Figura 1



La nube de Linked Data (los datasets y sus enlaces) en septiembre de 2010. Los datasets en verde de la izquierda, corresponden a datos de gobierno.

datos publicados por el gobierno para sus operaciones. En tercer lugar, la publicación de datos gubernamentales permite que la ciudadanía esté más informada sobre cuáles son las actividades del gobierno y cómo se realizan, aumentando la transparencia y el accountability de este último: por ejemplo, en Estados Unidos es posible ver qué funcionarios de la Casa Blanca han sido visitados, cuántas veces y por quién[1]. Finalmente, tecnologías como la Web permiten que el costo de publicar datos sea muy bajo: una vez que los datos han sido recolectados o generados y usados por el gobierno, el proceso de publicarlos es generalmente sencillo y simple.

En 2007 un grupo de expertos definió un conjunto de ocho principios que reflejan cómo los gobiernos debiesen publicar datos[2]: los datos deben ser completos, primarios, estar disponibles a tiempo, ser accesibles, fácilmente procesables por máquinas, no se debe discriminar a quienes lo soliciten, no deben estar en formatos propietarios y deben usar licencias abiertas. Existen por supuesto una serie de restricciones sobre qué cosas no pueden considerarse OGD: por ejemplo, es usualmente aceptado que la información personal de ciudadanos, así como datos que puedan afectar la seguridad nacional no deben ser publicados. A pesar de estas excepciones, se entiende como una buena práctica el que la opción por omisión sea publicar datos y el no hacerlo sea el caso particular.

LINKED DATA

Uno de los movimientos de mayor impacto en la Web Semántica es Linked Data[3], el cual consiste en una serie de principios para publicar datasets acerca de distintos temas, donde cada "cosa" (un auto, una persona, el día de ayer) tiene asignada una URI (similar a una dirección Web o URL). Como cada cosa es identificable por estas URIs, el siguiente paso es enlazar estos datasets, identificando qué URIs se refieren a la misma "cosa" o están relacionadas de alguna manera: de esta forma, es posible navegar por distintos datasets para obtener más información que la provista por una organización solamente. Por ejemplo, un médico puede buscar en DBpedia -una versión "semantificada" de Wikipedia- acerca de la proteína P53, encargada de la supresión de tumores y encontrará una descripción de ésta en varios idiomas, así como temas relacionados (oncología, proteínas, etc.). Luego, desde DBpedia es posible obtener las URIs con que esta proteína es descrita en otros datasets. Al acceder a estos nuevos datasets es posible encontrar qué enfermedades están asociadas a P53.

Es claro que las comunidades de OGD y Linked Data tienen mucho en común y pueden beneficiarse mutuamente, la primera usando Linked Data como plataforma, la segunda mostrando en OGD un caso de uso real. Actualmente, un porcentaje importante de la "nube" de Linked Data (los conjuntos de datos conectados) son datos de gobierno, como puede verse en la Figura 1.

Figura 2

Portal Data.gov del Gobierno estadounidense, permite buscar datasets relacionados con agricultura, defensa, medio ambiente y presupuestos, entre otros.

DESARROLLO DE OGD EN EL MUNDO

Las historias de OGD en Estados Unidos y en el Reino Unido son ilustrativas de cómo los gobiernos han adoptado distintos modelos de OGD y cuál ha sido su relación con Linked Data.

Estados Unidos: modelo Bottom-Up

(Disclaimer: estoy asociado con Tetherless World Constellation y he participado activamente en éste como parte del trabajo realizado por este laboratorio en colaboración con Data.gov).

En mayo de 2009, la administración del Presidente Barack Obama lanzó el sitio Data.gov[4] que fue la primera plataforma centralizada de publicación de datos en el mundo, construida por un gobierno. Comenzando con cerca de 40 datasets, actualmente provee sobre los 300.000, los cuales describen información relacionada con temas de energía, salud, migraciones, seguridad pública y muchos más. El uso de los datos ha sido aprovechado por una serie de aplicaciones, como DataMasher[5] (sitio especializado en crear *mashups*, es decir visualizaciones de cruza de datos) y Fly On Time[6] (sitio que permite saber cuántas son las demoras de vuelos en Estados Unidos), por nombrar algunas. Una de las prioridades de Data.gov era liberar la mayor cantidad de datos, bajo un proceso de publicación simple, por lo que se dio flexibilidad a los funcionarios de gobierno en cuanto a los mecanismos de publicación: es así que los datos han sido publicados principalmente como archivos XML, Excel, Comma-Separated Values (CSV), Really Simple Syndication (RSS), Keyhole Markup Language (KML o KMZ) y archivos Shapefile (SHP).

Otra medida tomada para simplificar el proceso de publicación fue apuntar a los datos localizados en los servidores de los organismos gubernamentales correspondientes, en vez de replicarlos en Data.gov; de esta forma se evitan problemas técnicos y se puede reusar buena parte de la infraestructura existente (por ejemplo, servicios que proveen feeds RSS). Desde hace algún tiempo, Data.gov (Figura 2) ha comenzando a publicar datos en RDF (Resource Description Framework), el lenguaje para datos en la Web Semántica. Este trabajo se ha hecho en conjunto con Tetherless World Constellation y ha implicado dos procesos paralelos: por un lado la conversión textual de los datos de manera automática, donde estos se extraen desde las tablas Excel y archivos CSV y se aplica una transformación genérica para generar RDF. El segundo proceso consiste en la publicación de datos mejorados, curados manualmente, donde se busca una representación más fidedigna de lo que los datos representan en el mundo real, que a la estructura de la tabla desde la que fueron sacados. Por ejemplo: la conversión automática de una tabla con nombre, apellido y dirección de una persona considerará los tres valores asociados a la misma entidad (la fila de la tabla); una versión mejorada considerará qué nombre y apellido pertenecen a una persona, mientras que la dirección está asociada a un lugar, el cual está relacionado con la persona, como se puede ver en la Figura 3.

Reino Unido: modelo Top-Down

En enero de 2010, el Gobierno del Reino Unido lanzó Data.gov.uk[7] (Figura 4). El enfoque británico fue diferente: se usó tecnología semántica y Linked Data desde un principio, por lo que en muchos casos (no todos) los datos están disponibles en RDF así como en su contraparte en formato CSV. Por ejemplo, cada escuela en el Reino Unido tiene una URI (por ejemplo, http://education.data.gov.uk/id/school/103335). El uso de Linked Data permite que al acceder a esta URI (sea posible obtener información relevante para la escuela: al usar un navegador como Firefox o Chrome obtenemos un documento HTML, pero también es posible escribir programas que vean los datos "puros" en RDF usando esta URI, los cuáles serán más fáciles de procesar que extraerlos desde el HTML.

Asimismo, el Gobierno británico dispuso de SPARQL endpoints (servicios Web donde es posible ejecutar consultas en SPARQL, el equivalente a SQL para datos semánticos) con información sobre distintas áreas (educación, transporte, etc.), de manera

que en muchos casos no es necesario descargar la información, sino que es posible consultarla directamente en los servidores del Gobierno.

CATÁLOGOS DE OGD

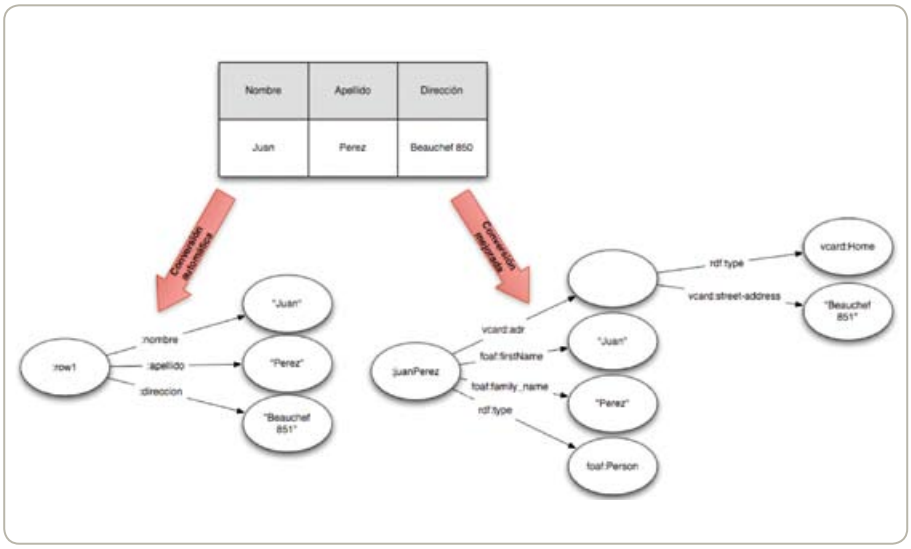
Como una forma de facilitar el acceso a esta gran cantidad de datos, existen varios esfuerzos por crear "metacatálogos" donde sea fácil buscar datasets disponibles en distintos portales. Así, por ejemplo, la Comunidad Europea ha trabajado en los últimos años para disponer de un portal centralizado que liste los datos de los países que la componen, tanto a nivel local, regional, así como nacional. Uno de los problemas es que al haber cientos de catálogos, no es fácil para los usuarios encontrar los datos que buscan, de manera que han creado PublicData.eu[8], el cual permite buscar en diversos portales de la Comunidad Europea. De esta forma, no se intenta replicar el trabajo hecho por otras organizaciones gubernamentales, sino agregarlo para facilitar la búsqueda por parte de los usuarios.

A nivel internacional, Tetherless World Constellation ha creado un catálogo de fuentes de datos de gobierno de diversos países y organizaciones internacionales, el cual se puede explorar seleccionando diversos criterios como país de origen y temas relacionados, entre otros[9]. Un esfuerzo similar ha realizado la fundación CTIC, la cual también provee un navegador[10] para buscar catálogos de datos por país y tipo, como se puede ver en la Figura 5.

DESAFÍOS

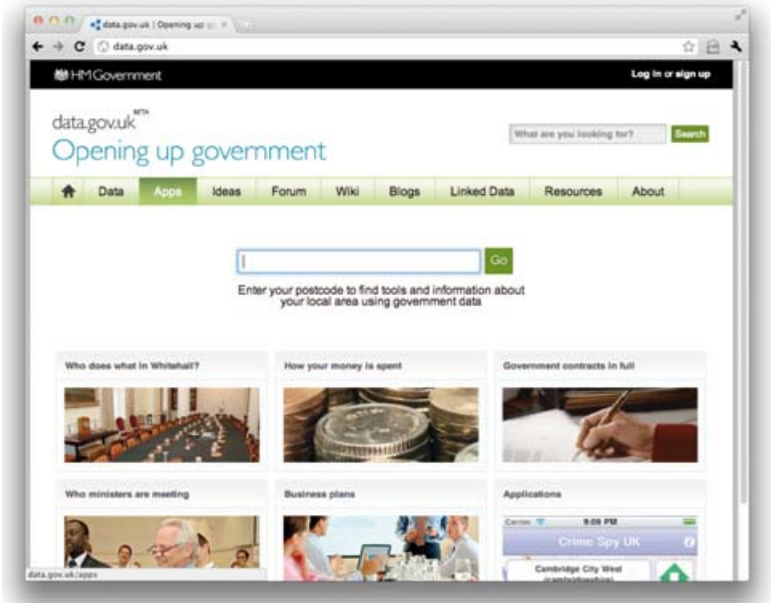
Se puede decir que OGD presenta desafíos en varios frentes, pero por brevedad sólo mencionaré los que parecen más relevantes. En primer lugar, es necesario un fuerte apoyo político y considerar a OGD como una política fundamental para mejorar la transparencia de un gobierno. Sin una valoración desde el mundo político, cualquier esfuerzo se va a quedar sólo en buenas intenciones. Más aún, este apoyo debe verse reflejado en una asignación de recursos, ya que como cualquier otra política pública, implementar OGD requiere tiempo y dinero para que los encargados lo puedan llevar a cabo. En segundo lugar, la cantidad, variedad y distribución de datos disponibles implican que se requiere especial preparación por parte de los organismos públicos: la experiencia en distintos países muestra que es necesario capacitar a quienes serán los encargados de OGD en cada órgano del Estado, lo que toma tiempo. En tercer lugar está el asunto de la calidad. Es claro que no todos los datos son igual de "buenos" en términos del "ruido" que poseen, cuán confiables son, cómo son representados, etc. Para ayudar a resolver esto, es necesario establecer una serie de

Figura 3



El proceso de conversión desde una tabla a RDF puede ser automático, llevando a una representación más cercana a la tabla (a la izquierda) o una conversión mejorada, reusando vocabularios (en este caso FOAF y vcard) y más cercana a lo que los datos describen (grafo a la derecha).

Figura 4



Portal Data.gov.uk del Gobierno del Reino Unido. En la versión actual se facilita la búsqueda de datos y aplicaciones relevantes para localidades específicas, basado en el código postal.

Figura 5

Países que poseen catálogos de datos públicos, según la Fundación CTIC.

métricas que ayuden a los consumidores de estos datos. Finalmente, quizás las preguntas más importantes que debiésemos tratar de resolver son: ¿cómo hacemos para que los ciudadanos comunes y corrientes puedan sacar el máximo provecho de estos datos sin tener que convertirse en hackers?, ¿qué tipo de servicios debiesen ofrecer los gobiernos para aumentar la participación ciudadana en las iniciativas de OGD?

Con todo lo anterior, queda la pregunta sobre cómo poder replicar estas iniciativas en otros países e instituciones. Por una parte, la experiencia muestra que no es necesario centralizar todos los datos, sino centralizar las búsquedas: los usuarios no tienen por qué cargar con la responsabilidad de saber dónde están los datos, sólo saber que pueden buscarlos en un solo sitio. Esto conlleva a que el repositorio debe coordinar con los diversos organismos proveedores de datos; lo anterior es posible usando vocabularios para describir catálogos de datos, tales como dcat[11] para comunicar qué datasets están disponibles. Asimismo, es importante publicar los datos en la mayor variedad de formatos posible, de manera de llegar a diferentes audiencias y disminuir las barreras para la creación de aplicaciones. Para lograr esto es recomendable tener un modelo de datos flexible desde el cual sea posible traducir y exportar a diferentes formatos. Es aquí donde RDF aparece como una excelente alternativa: convertir desde RDF a otros formatos resulta más fácil que desde, por ejemplo, CSV o Excel. Por otro lado, una crítica importante que se le ha hecho a Data.gov es la falta de recursos para mantener una comunidad de hackers y desarrolladores. El acceso a ejemplos de código, APIs (Application Programming Interface), tutoriales, documentación, etc. facilita el uso de los datos por parte de programadores, particularmente quienes desarrollan software en su tiempo libre. Otra crítica hecha a Data.gov (y en menor grado a Data.gov.uk) ha sido la calidad del sistema de búsquedas. Encontrar la información que se busca no resulta fácil, lo que desmotiva a los usuarios. Un esfuerzo para mejorar esto ha sido alpha.gov.uk, el cual ofrece sugerencias en una forma similar a lo que hace Google Instant[13]. Ésta y otras alternativas para mejorar las búsquedas pueden ser críticas para garantizar el éxito de un portal de OGD.

CONCLUSIONES

Este artículo ha hecho una breve revisión sobre qué es Open Government Data, su relación con Linked Data, así como ejemplos exitosos de la aplicación de estas tecnologías en gobiernos de distintas partes del mundo. Existen una serie de desafíos a la hora de implementar OGD: en general existe un conflicto natural entre la simplicidad de publicación y simplicidad de consumo de los datos y cada gobierno ha buscado un camino diferente para lidiar con este problema. Más aún, hacer fácil para la ciudadanía el usar estos datos sigue siendo un problema abierto. Sin embargo, ya es posible ver beneficios en el uso de estos datos por parte de empresas y desarrolladores para creación de aplicaciones y servicios. Asimismo, OGD ha mostrado que es posible transparentar las actividades del gobierno, facilitando la detección de potenciales fraudes e ineficiencias en la gestión.

Todavía hay mucho camino por recorrer para aprovechar todo el potencial que ofrece Open Government Data, pero la tendencia en el mundo es que poco a poco los gobiernos van abriendo más sus datos para que la ciudadanía pueda hacer uso de ellos tanto a nivel nacional, regional como local. BITS

REFERENCIAS

[1] http://bit.ly/WHvisitors

[2] http://www.opengovdata.org/home/8principles

[3] http://linkeddata.org

[4] http://data.gov

[5] http://www.datamasher.org/

[6] http://flyontime.us/

[7] http://data.gov.uk

[8] http://publicdata.eu

[9] http://logd.tw.rpi.edu/demo/international_dataset_catalog_search

[10] http://datos.fundacionctic.org/sandbox/catalog/faceted/

[11] http://vocab.deri.ie/dcat

[12] http://drupal.org

[13] http://www.google.com/instant/

OPEN DATA

En camino hacia la Web Semántica: **experiencias de la Biblioteca del Congreso Nacional de Chile**

Biblioteca de Valparaíso. **Gentileza: Biblioteca del Congreso Nacional.**

BCN

Biblioteca del Congreso Nacional de Chile (BCN)[1]

La Biblioteca del Congreso Nacional, al servicio de los parlamentarios, y en estrecha coordinación con el Senado y la Cámara de Diputados, es un espacio de interacción social entre estos y la comunidad nacional. Aquí pueden reconocerse en su historia político social, informarse y compartir conocimiento acumulado. Además, permite la vinculación con articuladores del conocimiento nacional y mundial en los ámbitos social, político y legislativo.

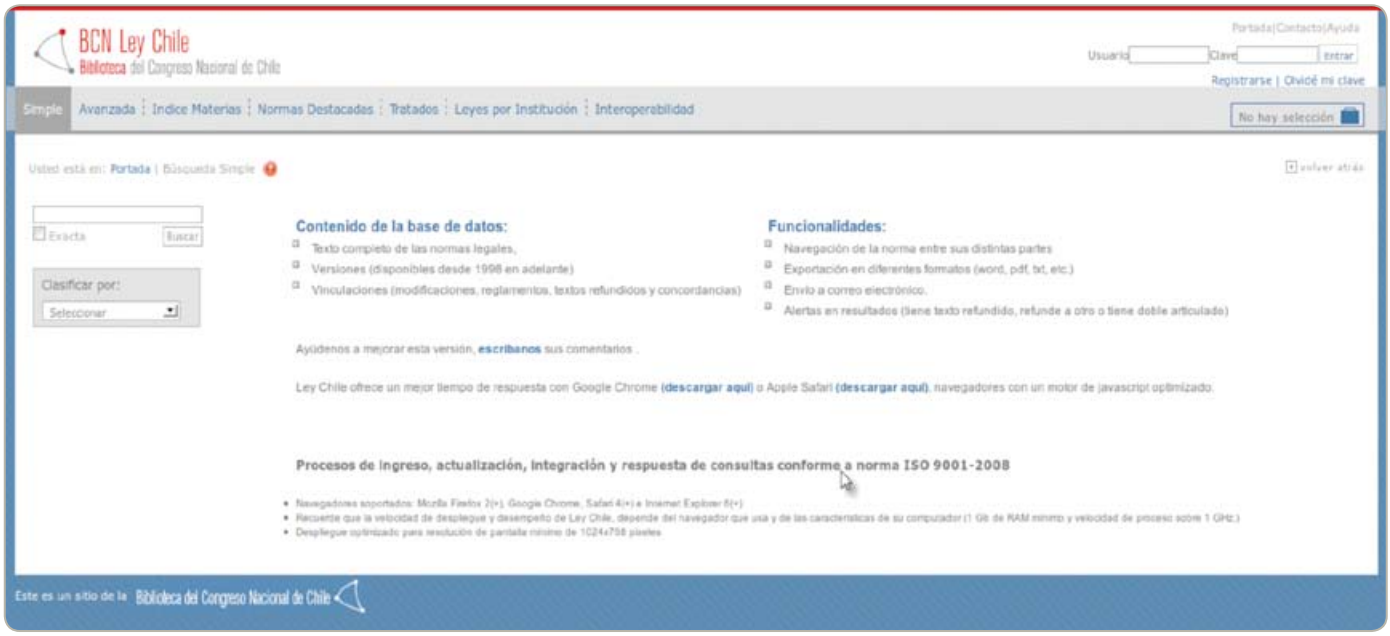
La Biblioteca del Congreso Nacional (BCN) adscribe al concepto de Open Government por considerar que se trata de una filosofía de trabajo útil para empoderar a los ciudadanos y otorgarles acceso y licencia de uso a los datos generados por entidades públicas, de tal manera que los puedan usar, almacenar, redistribuir e integrar con otras fuentes de datos. Esta apertura de la información se justifica tanto por favorecer la participación ciudadana, fortaleciendo la democracia, como por ser un motor de innovación al permitir la creación de nuevas industrias con estos datos.

Hoy en día el concepto de Open Government se entrelaza con los conceptos de Open Data y Linked Data[2]. Entendemos que mientras el concepto de "Open Data" se orienta a que los datos deben ser asequibles a todos en forma libre y sin restricciones, "Linked Data" es una forma de publicar los datos de manera tal que se facilite la interrelación entre las distintas fuentes de datos.

El año 2010 la World Wide Web Foundation[3] a través del Centro de Tecnología, CTIC, realizó un estudio acerca de la aplicabilidad y potencial de una iniciativa de Open Government Data (OGD) en Chile, Ghana y Turquía. El reporte [1] indica que Chile presenta condiciones propicias para la liberación de la información pública en términos de disposición para OGD, recomendando su inicio dada la existencia de conocimiento, personas, tecnología y voluntad.

1 Artículo escrito por: Christian Sifaqui, Eridan Otto, Felipe Almazán y Daniel Hernández (asesor externo).
2 Linked Data - Connect Distributed Data across the Web. www.linkeddata.org
3 World Wide Web Foundation. http://www.webfoundation.org/

Figura 1



Página principal de www.leychile.cl

Así como el informe lo indica, la BCN cree que los organismos estatales en Chile están preparados para entrar de lleno a OGD. En esta línea, la Biblioteca ha llevado a cabo las siguientes acciones:

- Portales/Sitios de libre acceso: toda la información que la BCN pone a disposición en sus portales (www.bcn.cl, www.leychile.cl, etc.) es completa, confiable, distribuible, reusable, se basa en el concepto de patrimonio cultural común y permite la interoperabilidad. Esto ha llevado consigo un incremento de las visitas de sus portales en forma constante, por ejemplo, el año pasado se contabilizaron más de 10,5 millones de visitas anuales y durante estos primeros siete meses ya hay un incremento del 19% con respecto al año 2010.
- Marcaje: la BCN coloca en los objetos digitales marcas y atributos semánticos para obtener resultados más precisos y relacionados entre sí. Este proceso de "semantizado de la información" ha permitido que los buscadores actuales accedan en forma precisa a la información disponible en nuestros portales.
- Widgets/Gadgets: la BCN entrega en sus portales www.bcn.cl y www.leychile.cl un conjunto de aplicaciones computacionales en plataforma Web como una forma de entregar acceso automático, distribuido y sencillo.
- Web Semántica: como punto de partida en el uso de las tecnologías que sustentan este concepto, la BCN liberó el sitio http://datos.bcn.cl donde se ofrecen datasets con ontologías públicas para facilitar el análisis computacional automático y se publicarán datos en el modelo de datos enlazados (RDF), algunos de ellos pueden ser accedidos con el lenguaje de consulta SPARQL. Hoy, datos.bcn.cl incorpora los datos provenientes del portal Ley Chile y progresivamente se irán incorporando otros datasets, como el de las Reseñas Biográficas de Parlamentarios. Junto con la publicación centralizada en datos.bcn.cl, se está trabajando en el marcado de contenidos con RDFa en varios de los portales administrados por la BCN, lo que facilitará la publicación de datos de manera distribuida y su posterior integración con datos.bcn.cl.

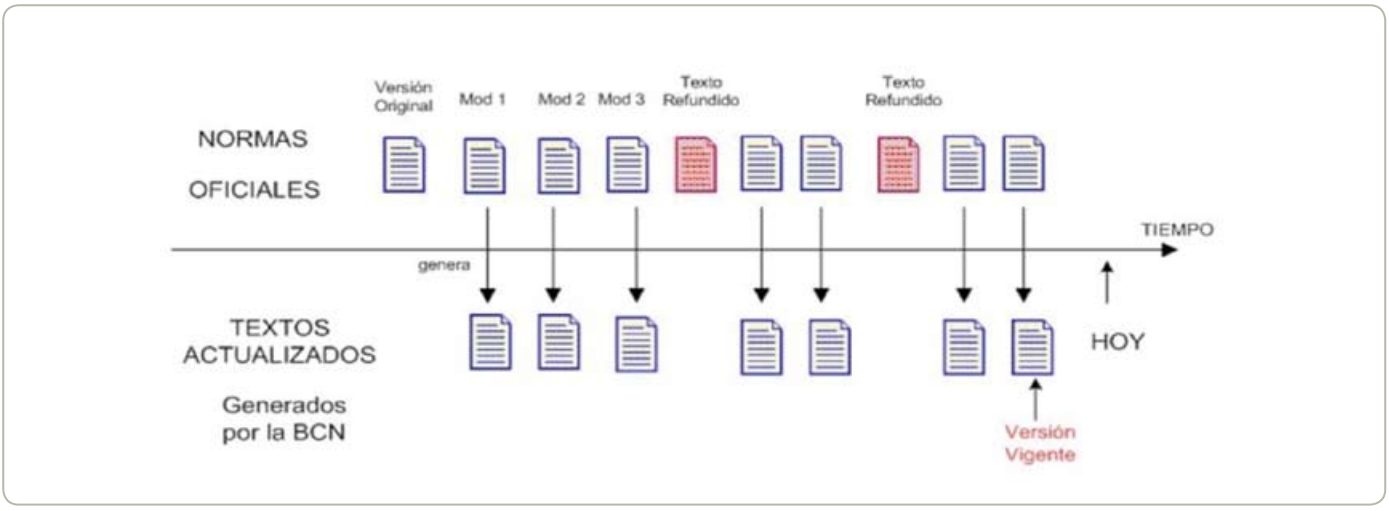
En este artículo se describen algunos proyectos realizados por la BCN que se enmarcan en torno a la iniciativa OGD, lo anterior ha sido un proceso natural realizado por la institución, ya que la naturaleza de sus servicios hace que esté permanentemente incrementando semánticamente la información que ella administra y pone a disposición de la ciudadanía.

ESTRUCTURACIÓN DE LOS CONTENIDOS CON XML: CASO LEYCHILE

Se define la legislación como el conjunto de normas positivas (leyes, decretos, resoluciones, etc.) que conforman el ordenamiento jurídico nacional.

La legislación es información que proviene del sector público, es generada por órganos del Estado y financiada con recursos públicos. Asimismo es de interés público, ya que opera en temas de utilidad general y afecta la vida de los ciudadanos en particular. Esta información tiene un alto valor público, ya que genera una experiencia en los

Figura 2



Modelo de "vida" de una norma.

ciudadanos que es considerada valiosa por ellos. Dicho lo anterior, podemos indicar que la legislación es información pública y por ende debe ser de dominio público.

Se desprende en forma natural el concepto de "mecanismo de concreción de principio de seguridad", que es la fundamentada expectativa que tienen los ciudadanos de que la ley vigente se cumpla. Para cumplir con esta expectativa, podemos reconocer dos enfoques: el concepto de "seguridad jurídica" (certidumbre fundada y garantizada que la norma será cumplida) y el concepto de "certeza jurídica" (perceptibilidad de la norma jurídica y la certidumbre de su contenido). Para satisfacer ambos enfoques, los países ofrecen un mecanismo de publicidad de la ley, conocido como Diario o Gaceta Oficial.

En el caso particular de Chile, tres artículos del Código Civil hacen referencia a este mecanismo de publicidad y de los enfoques mencionados, a saber:

- *Art. 7º. La publicación de la ley se hará mediante su inserción en el Diario Oficial, y desde la fecha de éste se entenderá conocida de todos y será obligatoria.*
- *Art. 8º. Nadie podrá alegar ignorancia de la ley después que ésta haya entrado en vigencia.*
- *Art. 706° [...] el error en materia de derecho constituye una presunción de mala fe, que no admite prueba en contrario.*

Pero los artículos mencionados nos conducen a lo que se conoce como "ficción legal del conocimiento", ya que en Chile el acceso al Diario Oficial es pagado, se publican las normas modificatorias en vez de los textos vigentes y el acceso a la normativa de períodos anteriores es dificultosa.

Como una forma de solucionar la "certeza jurídica" para el Congreso Nacional de Chile y también para los ciudadanos, la BCN en los años cincuenta inició mediante un sistema de fichas una recopilación de las referencias de las vinculaciones de las normas y clasificó las normas bajo materias.

En los años setenta este sistema de fichas fue reemplazado por un sistema STAIRS, que permitió automatizar estas fichas y sus anotaciones. Posteriormente, a mediados de los ochenta, fue reemplazado por un sistema cliente-servidor basado en BASIS PLUS, que permitía reconstruir en línea los textos de las normas.

En el año 2008 se libera el sistema Web LeyChile (ver Figura 1), el cual contiene todas las normas a texto completo, sus versiones (disponibles desde 1998) así como las vinculaciones (modificaciones, reglamentos, textos refundidos y concordancias). Este sistema ofrece una caja de búsqueda y al mismo tiempo servicios Web que ofrecen el texto en formato XML, también proporciona servicios complementarios y aplicaciones como widgets y gadgets para hacer más fácil el consumo y uso de la información legal, almacenada en esta base de datos.

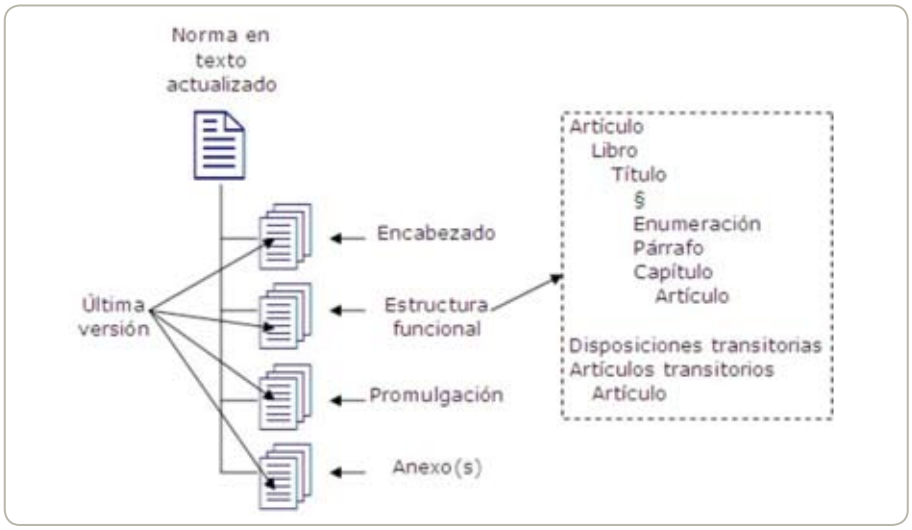
Hoy en día, este sistema tiene en promedio unas 14.000 visitas diarias, alcanzando picos de 18.000 visitas, lo cual consideramos muy alto tomando en cuenta el tipo de contenidos y la cantidad de 7,3 millones de usuarios de Internet en Chile [2].

Detalles de la implementación

LeyChile fue concebido con una arquitectura básica de tres capas:

- **Datos:** representación y almacenamiento XML nativo (en una base de datos híbrida) para las normas y sus vinculaciones. Las normas se encuentran indexadas para responder consultas a texto completo, tanto a nivel de norma como un todo o de sus partes. Existen servicios básicos como entrega de la norma completa para su procesamiento hacia las capas superiores e imponer algunas reglas de negocio.

Figura 3

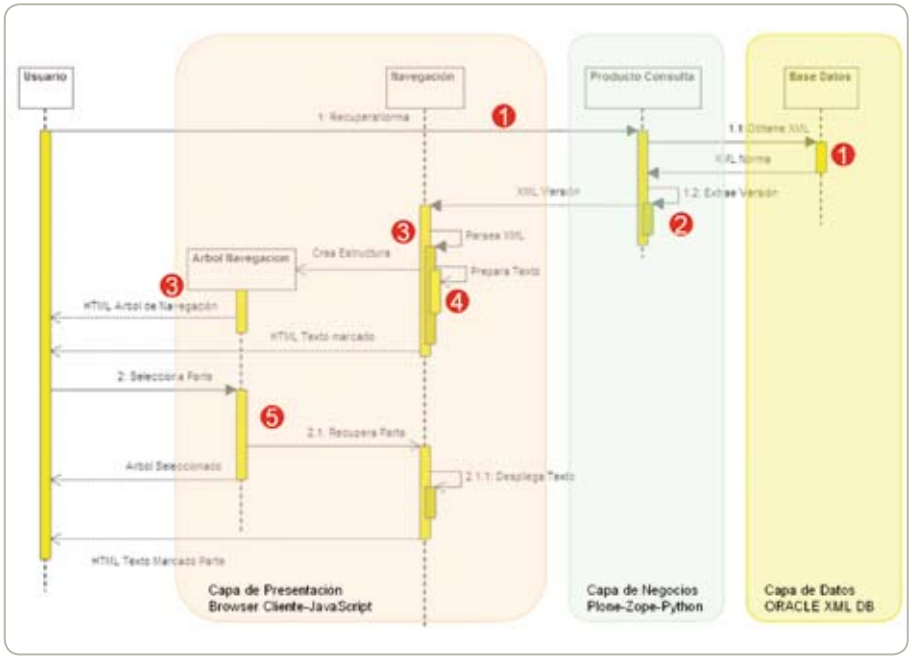


Modelo de la estructura de una norma legal.

- **Negocios:** aplica las transformaciones necesarias a la norma, como el cálculo de las partes asociadas a una versión, generando XMLs para ser procesados por la capa de presentación. En esta capa el servidor de aplicaciones Web (ZOPE-Plone) arma las páginas dinámicas del sistema y habilita el procesamiento AJAX tanto para las páginas HTML como para los servicios Web.
- **Presentación:** una parte importante del procesamiento visual y las características interactivas de la navegación de la norma son distribuidas a los clientes por medio de código JavaScript.

Un concepto básico que fue tomado en cuenta es que se considera que la norma no es un objeto estático, por el contrario, durante su ciclo de vida desde que se publica en el Diario Oficial hasta que eventualmente es derogada o refundida, sufre modificaciones. Tal como se muestra en la Figura 2, una norma sólo puede ser modificada por otra norma (modificatoria), lo que genera una nueva versión de la misma. El modelo XML de la norma permite la modificación sólo a las partes (encabezamiento, estructura funcional, promulgación, anexos, ver Figura 3) que son afectadas por la modificatoria. De esta manera LeyChile construye dinámicamente el texto completo de una versión (versión vigente, intermedia u original).

Figura 4



Mecanismos para navegar en la norma en forma interactiva.

Otro detalle a considerar es que en algunos casos el Poder Ejecutivo genera una versión oficial de la norma, denominada texto refundido, con la finalidad de sistematizar, coordinar y ordenar el contenido de una norma que ha tenido una cantidad importante de modificaciones. LeyChile modela esto mediante un enlace (vinculación) entre el texto refundido y la norma que le dio origen.

Cabe hacer notar que todo el modelo del documento normativo es en su fase conclusiva (promulgado) sin ocuparse de todo el íter legislativo.

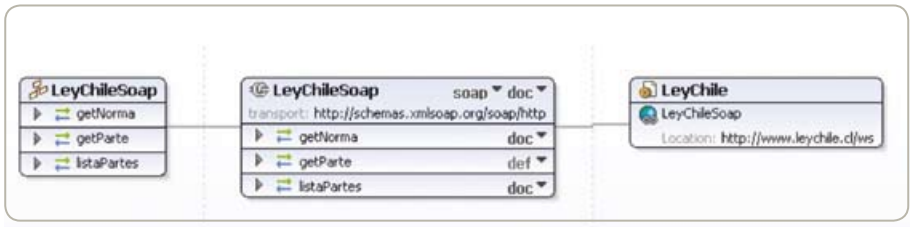
En base a los detalles anteriormente expuestos, el esquema XML de LeyChile, distingue tres capas o niveles de marcado:

- **Texto:** versiones, hipervínculos, referencias, notas.
- **Estructura:** organización jerárquica de las partes de una norma (ver Figura 3).
- **Metadata:** conocimiento adicional al documento formal, por ejemplo, identificación de la norma, materias, términos libres, etc.

Esta estrategia permite a diferentes actores de la organización "enriquecer" el texto legal, corregir y aumentar estas marcas si es necesario. LeyChile consta de un módulo de producción que gestiona un flujo de trabajo de analistas especializados que van completando el marcado de la norma.

El mecanismo de recuperación de una norma desde la base de datos XML y su navegación interactiva desde un browser ha sido optimizado para reducir la carga de procesamiento de los servidores de datos y de aplicación, por lo que el procesamiento principal se realiza en el browser de cada cliente, el cual parsea la estructura XML entregada, haciendo fuerte uso de las capacidades asincrónicas de JavaScript. La Figura 4 ilustra las interacciones que se

Figura 5



SOAP, disponible en http://www.leychile.cl/ws/LeyChile.wsdl

desencadenan cuando el usuario selecciona una parte de la norma:

1. **Recuperación:** el usuario selecciona una norma para navegar, la capa de negocios realiza una petición a la base de datos, la cual extrae el XML de la norma completa.
2. **Extracción versión:** la capa de negocios extrae el XML de la versión requerida de la norma.
3. **Parseo:** el browser del cliente parsea el XML recibido, creando una estructura de datos que representa la organización jerárquica de la norma, la cual se muestra gráficamente en forma de árbol de navegación.
4. **Despliegue:** se genera en forma asincrónica un HTML dinámico con el texto de la norma completa en conjunto con elementos gráficos, como las notas y los metadatos asociados.
5. **Navegación:** en la medida que el usuario selecciona las distintas partes de la norma que requiere en detalle, se realiza el mismo procesamiento descrito en el punto cuatro, pero a nivel de una parte.

La arquitectura del sistema se diseñó para enfrentar una serie de desafíos respecto al rendimiento y tiempo de respuesta:

- Se supuso una gran cantidad de visitas que producirían una fuerte carga en los servidores de datos y de aplicación.
- Se definió el hospedar el mayor tiempo de respuesta en el navegador del cliente debido al proceso de parseo, despliegue y navegación interactiva de las normas, en especial normas de gran tamaño como los códigos legales.

Se llegó a la solución actual mediante procesos iterativos de optimización de los algoritmos básicos de procesamiento del XML de la norma, basados en uso de capacidades asincrónicas de JavaScript y el desarrollo de varios servicios de caché de datos propios (normas y partes de normas preformateadas, PDF, etc.), mantenidos en un servidor NFS compartido. Los datos que se recuperan siempre están actualizados mediante mecanismos de limpieza de las partes preformateadas, cada vez que se producen cambios sobre las mismas.

Ley Chile ofrece interoperabilidad con otros sistemas, tanto internos como externos, mediante Web Services. Los Web Services hacen uso de mensajería en lenguajes basados en XML. Algunos servicios utilizan el mecanismo REST y otros el protocolo SOAP. Utilizar este estándar de integración permite independizar la interacción de la tecnología específica utilizada al interior de cada aplicación de la BCN o de las tecnologías utilizadas en cada institución.

La Figura 5 muestra la definición de los servicios SOAP. Por ejemplo, getNorma, permite a cualquier sistema en la Web consumir el XML de intercambio[4] de una norma en particular.

La mayoría de los servicios son invocados mediante un llamado HTTP y entregando un XML de respuesta, por ejemplo: últimas leyes publicadas[5], proyectos de ley despachados por el Congreso hacia el Ejecutivo o hacia el Tribunal Constitucional, metadatos de una norma[6] entre muchos otros.

SEMANTIZACIÓN DE LOS CONTENIDOS CON RDFa

Reseñas biográficas de parlamentarios

Desde la fundación del Congreso Nacional de Chile en 1811, hacia 2011 han desempeñado un cargo de representación ciudadana como parlamentarios (diputado o senador) más de 3.800 personas. La Biblioteca del Congreso mantiene actualizada una reseña biográfica para cada una de ellas.

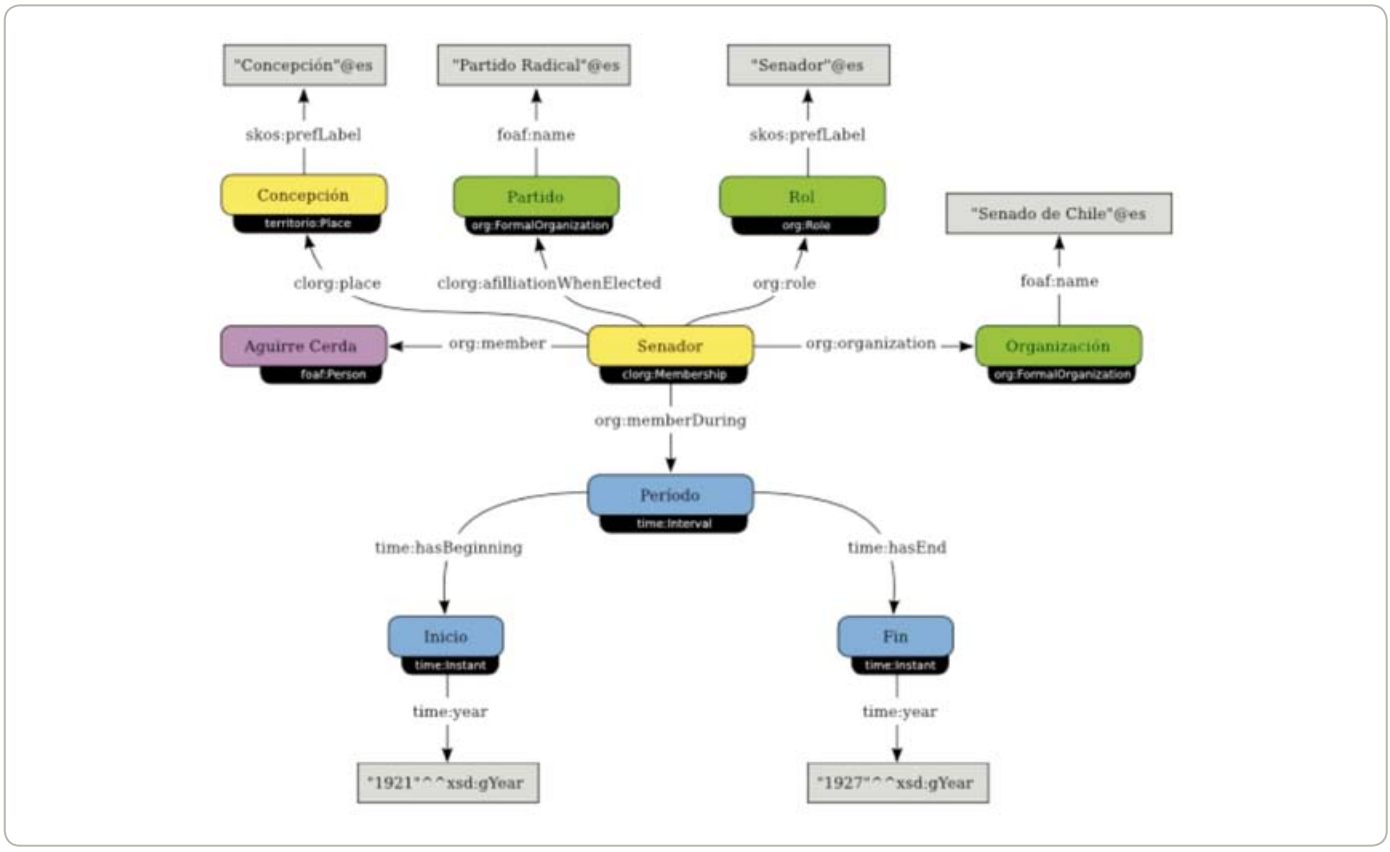
El trabajo de investigación, recopilación y confección de las reseñas se inició hacia 2001. En 2003 se logró completar una breve reseña biográfica para todos los parlamentarios y se liberó un sistema de

Figura 6

Reseña Biográfica Parlamentaria en portal de Historia Política.

4 Esquema aceptado como el modelo estándar de la norma chilena para intercambio, por la administradora de esquemas y metadatos (inscripción 354) http://www.aem.gob.cl/index.html
5 http://www.leychile.cl/Consulta/obtxml?opt=3&cantidad=5
6 http://www.leychile.cl/Consulta/obtxml?opt=4546&idNorma=206396

Figura 7



Esquema de vocabulario asociado a la trayectoria parlamentaria.

base de datos para consultas a través del sitio Web de la BCN. En los años siguientes se profundizó este trabajo mediante la extensión y completitud de las reseñas y la asociación de material referencial como fotografías, artículos de prensa y enlaces, entre otros. En 2009 se liberó una nueva versión de este sistema en plataforma wiki[7].

Con motivo del Bicentenario del Congreso Nacional, la BCN integró el sistema de Reseñas Biográficas Parlamentarias dentro del portal "Historia Política Legislativa del Congreso Nacional de Chile"[8], el cual fue liberado en julio de 2011 (Figura 6).

Este portal ofrece la opción de consultar directamente por el nombre de un parlamentario o acceder a listados alfabéticos de todos los parlamentarios o sólo de aquellos que se encuentran actualmente en ejercicio. Asimismo, las reseñas se presentan agrupadas en seis períodos relevantes de la historia de Chile.

Estructura de la Reseña Biográfica

Cada reseña está compuesta de dos partes fundamentales:

- Un relato biográfico que recopila sus orígenes familiares, estudios y trayectoria profesional, su labor legislativa, además de información relacionada como artículos de prensa, material digital y fuentes referenciales.
- Una ficha resumen que destaca una fotografía, la trayectoria parlamentaria, cargos públicos, antecedentes personales y enlaces relacionados a sitios Web:
- La trayectoria parlamentaria indica por cada período de representación el cargo desempeñado, el partido político en el cual militaba al momento de ser electo y el nombre del parlamentario que lo precedía, el cual está enlazado a su respectiva reseña. En el caso de los parlamentarios actuales, se enlazó la zona geográfica que representa (división político-electoral[9]), al Sistema de Información Territorial (ver Figura 7).
- Los antecedentes personales muestran el nombre completo del parlamentario, la fecha de nacimiento, lugar de nacimiento, fecha de fallecimiento, lugar de fallecimiento y profesión. En el caso de los parlamentarios actuales, se incluyen enlaces a páginas como Facebook, Twitter, Web personal, así como sus fichas en los portales de la Cámara de Diputados y el Senado, respectivamente.
- Los cargos públicos, que indican el cargo desempeñado (generalmente ministros de Estado), el período asignado, el nombre del Presidente de la República para el cual desempeñó este rol y los nombres de quienes lo precedieron y sucedieron en el cargo.

7 Reseñas parlamentarias. http://biografias.bcn.cl
8 Portal Historia Política. http://historiapolitica.bcn.cl
9 División Político-Electoral. http://siit2.bcn.cl/divisionelectoral/index.htm

Figura 8

Portal de acceso a http://datos.bcn.cl

Solución implementada

Se consideraron cinco etapas: definir una ruta única para cada reseña parlamentaria; investigar y seleccionar y/o construir vocabularios; marcar documentos con etiquetado RDFa; definir consultas y mecanismos de recuperación de información, y publicación de vocabularios en portal de Linked Data BCN.

Los vocabularios utilizados son: Dublin Core, Friend of a Friend (FOAF)[10], Licencia Creative Commons, Open Provenance Model Vocabulary (OPMV), Biographical Information (BIO)[11], Simple Knowledge Organization System (SKOS), Time (TIME) y Core Organization Ontology (ORG).

Marcaje de Contenido RDFa

El marcado de la wiki con RDFa se basó en las plantillas que ya eran utilizadas para generar los cuadros de datos que acompañan la información narrada en las páginas de la wiki. Una modificación de estas plantillas permitió reutilizar los datos marcados con la sintaxis de MediaWiki12, generando XHTML+RDFa en vez de HTML plano como se venía haciendo. El sistema de plantillas de MediaWiki incluye un lenguaje funcional que permite definir funciones en páginas del wiki cuyos parámetros son los datos a representar. Al encontrar llamadas a estas funciones el sistema las reemplaza por el resultado de evaluarlas. La evaluación consiste en reemplazar los valores de los parámetros en la plantilla que corresponde a la función, que es también una página del wiki. Dado que las plantillas son también páginas wiki, en ellas se pueden llamar otras funciones.

Acceso y consulta de los datos

Al incorporarse el marcado en las páginas Web, los datos quedan automáticamente accesibles a quien desee extraer los triples, procesarlos o integrarlos con otras fuentes de datos. Además, para facilitar el uso e integración de datos publicados, la BCN ha puesto un SPARQL endpoint, donde publicará todos los triples divulgados en sus sitios. Actualmente el endpoint sólo cuenta con los datos provenientes del portal Ley Chile, pero a medida que se consoliden, se irán incorporando nuevos datasets, como el desarrollado para Reseñas Biográficas.

WEB SEMÁNTICA: CASO DATOS.BCN.CL

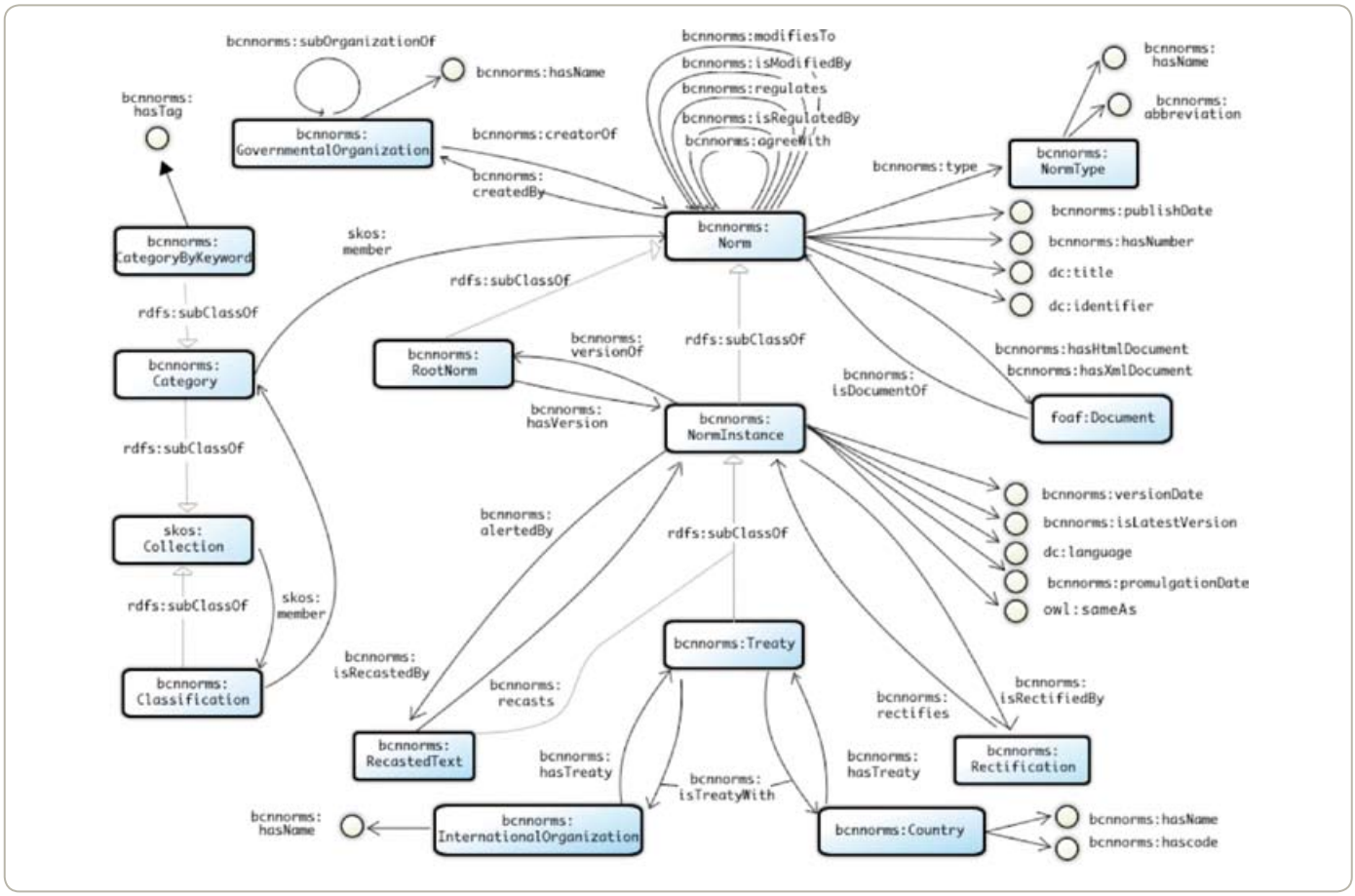
Este proyecto tiene como objetivo entregar a los ciudadanos acceso a nuestras fuentes de datos como Linked Open Data. Este proyecto está operativo desde fines de mayo de 2011 y es la primera iniciativa de la BCN con relación a la publicación de Linked Data (Figura 8).

Creemos que ofreciendo esta nueva capa de servicios, todo el sistema será un modelo de referencia en términos de entidades y reglas, publicado mediante ontologías e incluyendo la posibilidad de ejecutar consultas complejas con información de las normas legales a través de un endpoint SPARQL, permitiendo así acceder a resultados en variados formatos de salida, tales como RDF/XML, JSON, HTML+RDFa o N3.

En una primera etapa se desarrolló el proceso de contextualización bajo un dominio muy puntual, el de normas en el contexto legislativo de ofrecer una nueva capa de servicios a LeyChile (mencionado en la sección "Estructuración de los contenidos con XML: caso LeyChile" de este artículo). Para ello, se redactó un documento donde se describieron los tres elementos principales del contexto: qué datos se van a entregar, la forma de entregarlos y quién va a consumirlos. En orden a las interrogantes anteriores y de manera muy sintetizada, los datos a entregar son normas y sus relaciones, sin considerar en esta primera etapa la estructura interna de una norma, la forma de entregar los datos es a través de un grafo RDF sobre HTTP y un Endpoint SPARQL, y por último, quienes van a consumir los datos son aplicaciones de visualización de datos de la misma Biblioteca y aplicaciones orientadas a la consulta de leyes que puedan ser implementadas tanto por administraciones públicas como por

10 Friend of a Friend (FOAF) foaf="http://xmlns.com/foaf/0.1/"
11 Biographical Information: bio="http://purl.org/vocab/bio/0.1/"

Figura 9



Ontología de Normas.

la comunidad. Posteriormente se definió una ontología (ver Figura 9) y un espacio de nombres para la ontología de normas en el contexto particular de la realidad nacional.

Se ha considerado una estructura extensible de la ontología a otros dominios tales como congreso, educación, salud u otros. Esta ontología ha sido escrita usando RDF Schema y OWL, permitiendo así la aplicación de inferencias al grafo RDF. Otra característica importante de esta ontología es que ha sido compuesta usando ontologías previas y datasets como SKOS, Dublin Core, FOAF, Geonames, Organization y DBPedia. Usando las dos últimas mencionadas, fue posible enlazar datos del grafo de normas legales a conjuntos de datos externos, específicamente respecto de tratados internacionales y países. Esta tarea no fue trivial porque requirió un intenso trabajo manual. Finalmente, la ontología se almacenó en el RDF store para así permitir inferencias como las ya publicadas usando archivos de textos en RDF/XML y sintaxis N3, mientras que su documentación fue publicada en castellano e inglés.

Una vez estructurada la ontología, se modeló el grafo de salida RDF. En la práctica, se definió un esquema URI con todos los patrones URI posibles que podrían ser consultados de forma válida. La Figura 10 muestra un ejemplo de un patrón URI, el cual tomó en consideración el uso del estándar IFLA FRBR como una URI de las normas legales.

En términos generales, el grafo sigue un esquema jerárquico en cada uno de los recursos disponibles para consultas. Por otro lado, se modelaron algunas consultas (por ejemplo, obtener normas legales para fechas específicas). Así, para cada patrón URI se definió una salida RDF usando sintaxis N3. Finalmente se definieron los formatos de salida para los recursos. Para este proyecto, fueron definidos RDF/XML, JSON, Ntriples, N3 y HTML+RDFa.

Posteriormente se generó el proceso de transformaciones y carga de datos. Para este fin, se construyó un servicio de actualización en Java usando la API Kettle para el proceso de carga, el proceso de actualización y el proceso de transformación. Así, usando el diseñador ETL se implementaron las diferentes transformaciones que generan las triplas RDF en sintaxis N3, para la carga inicial y para la actualización de las triplas (que usualmente sólo agregará nuevas triplas). Bajo estas condiciones el servicio de actualización ejecuta transformaciones y después carga las triplas en el RDF store.

En la siguiente fase, el grafo de salida RDF sobre HTTP fue implementado de acuerdo al modelo diseñado para ese propósito. Para su implementación se usó la herramienta WESO DESH, un front end de Linked Data,

Figura 10

```
cl/{type}/{organization}/{publish_date}/{number}/data.{ext}
```

Patrón de URI.

que se liberará próximamente como software libre. Finalmente, esta implementación de Linked Data fue certificada con validadores de Linked Data como Vapour de la Fundación CTIC y RDF/XML de W3C.

El proyecto cuenta con un portal Web de documentación y en forma constante se le agrega nueva documentación en castellano e inglés acerca de cómo usar la infraestructura Linked Data.

Cabe destacar que este proyecto fue desarrollado enteramente con software libre, de tal manera que un emprendimiento similar podría ser replicado sin pagar licencias de software. Este proyecto está finalizado y asequible bajo la URL http://datos.bcn.cl

Una particularidad de este proyecto es que dada la naturaleza de los datos, es normal encontrar discrepancias en la redacción o errores de tipeo, por lo que se tomó especial cuidado en el diseño del modelo para permitir editar manualmente los datos en el futuro. Por ejemplo, se puede mencionar la instancia Governmental Organization definida en la ontología, debido a las variaciones de nombres para el mismo recurso, se definieron diferentes clases.

Una completa descripción de la arquitectura definida para el desarrollo de este proyecto se encuentra en el artículo *"Towards an architecture and adoption process for Linked Data technologies in Open Government contexts – A case study for the Library of Congress."*

CONCLUSIONES

La BCN ha ido introduciendo los estándares de la Web Semántica en forma paulatina e iterativa mediante el desarrollo de aplicaciones prácticas. Esta estrategia ha permitido a la BCN recorrer los diferentes estados que conducen a la implementación de una arquitectura semántica:

- Identificación universal (URI) y un conjunto de caracteres universal (Unicode). Plataforma Web de la BCN.
- Formatos de representación e intercambio de documentos y metadatos.
  - XML: Ley Chile
  - RDFa: Reseñas biográficas
- Datos enlazados, endpoint.
  RDF, SPARQL: datos.bcn.cl
- Modelamiento semántico
  OWL: Ontología para Ley Chile

Estos proyectos han entregado la experiencia y fundamentos para que la BCN se mueva exitosamente en el camino hacia la Web Semántica.

> La BCN adscribe al concepto de Open Government por considerar que se trata de una filosofía de trabajo útil para empoderar a los ciudadanos y otorgarles acceso y licencia de uso a los datos generados por entidades públicas, de tal manera que los puedan usar, almacenar, redistribuir e integrar con otras fuentes de datos.

Actualmente se está desarrollando un proyecto para semantizar mediante RDFa nuestro portal de Transparencia, y se está haciendo un uso intensivo de las tecnologías de Web Semántica para un proyecto de un nuevo sistema de Historia de la Ley y Labor Parlamentaria. Asimismo respecto de la interoperabilidad se está analizando la aplicación de RIF para poder relacionar distintas bases jurídicas existentes en los diversos organismos del Estado.

La BCN también está dedicando esfuerzos para fortalecer su infraestructura de información. En la actualidad están en estudio los proyectos RDF Book Mashup[12] y Open Library[13].

Como indicáramos al inicio de este artículo, la BCN tiene en su ADN semantizar la información de su acervo, las tecnologías actuales usadas con esta perspectiva han permitido un desarrollo acorde para ofrecer nuevos productos y servicios al Congreso, a los ciudadanos y a todo el país. Creemos que esta experiencia se puede replicar en todas las organizaciones del Estado y lograr así una base para un Open Government real. BITS

REFERENCIAS

[1] Open Government Data. Feasibility Study in Chile. Carlos Iglesias, ed., 2011, http://public.webfoundation.org/2011/05/OGD_Chile.pdf

[2] Latin america's internet population grows 15 percent in past year to 112 million people, A. Fosk, March 2011, http://www.comscore.com/Press_Events/Press_Releases/2011/3/Latin_America_s_Internet_Population_Grows_15_Percent_in_Past_Year_to_112_Million_People

[3] Towards an architecture and adoption process for Linked Data technologies in Open Government contexts – A case study for the Library of Congress of Chile. Francisco Cifuentes, Christian Sifaqui and José Labra. Proceedings of the 7th International Conference on Semantic Systems, 2011.

12 RDF Book Mashup: Serving RDF descriptions of your books. http://www4.wiwiss.fu-berlin.de/bizer/bookmashup/
13 Open Library: One web page for every book. http://openlibrary.org/

Open Source Software: similitudes y diferencias con Open Data

Jens Hardings
Gerente adjunto de Spitec Ltda. (www.spitec.cl). Ingeniero Civil en Computación DCC Universidad de Chile; Doctor en Ciencias mención Computación, Universidad de Chile. Sus áreas de investigación han estado principalmente ligadas al FLOSS (Free / Libre / Open Source Software) o Código Abierto / Software Libre, Tecnologías de Información y Seguridad.
jens@hardings.cl

Lo primero que salta a la vista cuando se considera Open Data y la corriente de Open Source Software o Software Libre (en adelante le llamaré FLOSS para no perdernos en distinciones ético-filosóficas muy específicas), es que existe una filosofía en común. En ambos casos se busca poder reutilizar el esfuerzo para crear una obra intelectual, sea ésta software o datos y así lograr una mejor eficiencia en el uso de los recursos a nivel macro. Al mismo tiempo, se da acceso a todos, sin distinción del uso. Gracias a este acceso universal y a que no existen restricciones al uso y la copia de la obra intelectual o a obras derivadas de ella (esencialmente, modificaciones, o algún software/datos existentes a los cuales se agregan otros), existen muchas visiones diferentes y normalmente tiende a haber una convergencia en el uso, así como integraciones que se realizan con facilidad al procurar utilizar estándares accesibles a todos.

El tema del acceso universal puede sonar trivial, pero siempre existe la tentación de querer influir para bien en el mundo restringiendo el uso de software o de información a los fines "correctos". Sin embargo, cualquier restricción típicamente afecta más a quienes honestamente quieren utilizar el software o los datos para fines generalmente considerados como correctos, y no detienen demasiado a quienes tienen malas intenciones.

En cuanto a diferencias entre FLOSS y Open Data, podemos mencionar que en Open Data, al menos hasta el momento,

no existen estándares definidos, sino más bien principios o buenas prácticas. Y posiblemente no tiene demasiado sentido definir estándares muy estrictos respecto de modelos de datos y similares, dada la diversidad de modelamientos, realidades e interpretaciones. En ese caso, es útil considerar la dupla Open Data + FLOSS para poder procesar los datos provenientes de diversas fuentes de forma útil y productiva.

Otra diferencia sustancial es que las fuentes de datos en Open Data en general son gubernamentales o cuentan con algún tipo de financiamiento público. En cambio, si bien para el FLOSS esa línea se hace muy razonable, en la práctica los proyectos tienden a surgir de iniciativas privadas, y en muchos casos personales más que institucionales.

Una característica que parece ser propia de FLOSS más que de Open Data es el concepto de reciprocidad, cuya instancia más conocida es la cláusula de Copyleft en la licencia GPL. Esencialmente, ésta es una restricción que se impone a quien recibe una licencia para usar, modificar y redistribuir el software, de que no puede entregar o redistribuir ese software u obras derivadas de él bajo una modalidad de licenciamiento diferente a la cual él (o ella) recibió. O sea, si se redistribuye el software o algún software derivado de él bajo una licencia que no sea GPL, se pierde el derecho de uso original y con ello cualquier derecho de modificación, redistribución, etc. El objetivo de esta cláusula es perpetuar la libertad que entrega la GPL mediante el mecanismo de restringir las restricciones que se pueden imponer a un tercero, o imponer un "prohibido prohibir".

CONCEPTO DE AUTONOMÍA

Si bien una de las tentaciones al considerar el uso de FLOSS es asumir una ventaja de precio o más bien de costo, en la práctica esta ventaja, cuando efectivamente la hay, tiende a no ser tan relevante como se pensaría en primera instancia. Lo que es realmente una ventaja que resalta por sobre todas las demás, pero en general es subestimada, es tener mayor control sobre la plataforma tecnológica. El uso de estándares preferentemente abiertos y bien documentados es sólo el comienzo, porque dependiendo de la necesidad y de la envergadura del usuario, es posible realizar desde pequeñas adaptaciones, que se vuelven parte del proyecto original con lo cual la mantención futura no recae en el usuario, incluso es posible desde influir en el desarrollo futuro de un proyecto, hasta liderar ese desarrollo o uno alternativo si no hay consenso.

A ese mayor control le llamo autonomía, y es un símil al concepto de soberanía que existe a nivel de Estado. Mientras mayor autonomía o soberanía tenga un ente a través de tener el control de las herramientas esenciales que requiere para funcionar, menos le afectan desde decisiones externas -pasando por crisis de proveedores o socios estratégicos- hasta lisa y llanamente prácticas hostiles. Por otro lado, no tiene demasiado sentido intentar tener una autonomía del 100% tal como a un país no le conviene cerrar su economía para evitar incidencia extranjera y que no le afecten potenciales crisis mundiales. Un ejemplo concreto es el primer acercamiento que tuvo Venezuela al Software Libre, donde más que participar de una comunidad existente parecía que se buscaba realizar una comunidad completamente autónoma y separada al interior del país. Lo importante es lograr que el nivel de autonomía o dependencia de una empresa, tal como lo debiera ser el nivel de soberanía o dependencia económica de un país, sea consecuencia de una decisión y no un aspecto que se deje al azar. Por lo mismo, las decisiones de qué software utilizar en la gestión de la cual depende una organización, es una decisión estratégica importante porque genera un nivel de dependencia que por lo general no es considerado y que le corresponde a la alta dirección definir.

ROL DEL FLOSS HOY

Ya vimos que el FLOSS es un componente importante cuando una entidad busca aumentar su autonomía, y por sobre todo, es un referente en ese aspecto contra el cual poder comparar otras soluciones. Por lo mismo, es importante que se mantengan y se nutran los proyectos, para no perder esa oportunidad y ese referente. Basta recordar la época en la cual las prácticas comerciales de los dominadores del mercado obligaban a actualizaciones masivas solamente para poder seguir siendo compatibles con los demás, y el software libre ha jugado un rol fundamental en cambiar esa realidad.

Lo anterior se traduce en que el FLOSS actualmente cumple, aparte de los casos de nichos que no detallaremos aquí, al menos tres roles fundamentales en la industria TI:

1. Infraestructura base y commodity

La industria TI está muy enfocada en innovar a un ritmo muy acelerado, tanto así que quien no innova en esta industria no puede optar a ingresos interesantes. Como consecuencia, la infraestructura base y/o commodities son cada vez menos interesantes para la industria, que debe mostrar su valor agregado para seguir obteniendo ingresos y en algunos casos mantener una infraestructura base, con todos sus costos asociados, solamente para poder mantenerse en la pelea por las innovaciones de punta. De forma natural la infraestructura base y las aplicaciones que realmente son commodity debieran ser dominadas por soluciones FLOSS, ya sea porque un proyecto FLOSS se impone como el dominante en ese nicho, o porque un actor dominante en el mercado decide que le conviene dejar su solución disponible bajo un modelo FLOSS, que mantener su desarrollo bajo su exclusivo alero (y centro de costos), siendo por lo tanto candidato natural a permanecer ahora como solución FLOSS liderando ese nicho.

2. Cumplimiento de estándares: validación y "glue"

Este es un aspecto particularmente interesante considerando la temática Open Data, ya que el FLOSS siempre, y particularmente en todo el desarrollo de TCP/IP y todos los protocolos relacionados a Internet, ha sido muy apegado a estándares públicos y abiertos.

Las razones para que un proyecto FLOSS opte por utilizar estándares abiertos son bastante directas: en lugar de invertir esfuerzo en encontrar una solución a un problema, se opta por una solución existente que cumpla con cierto nivel de calidad y ojalá de validación, y tampoco existen razones estratégicas por las cuales realizar un desarrollo propio. Mucho menos hay justificaciones para mantener en secreto ciertas partes, dado que se diseña pensando en publicar todo el código. Por otra parte, muchos creadores de estándares realizan una implementación que luego publican bajo alguna licencia FLOSS, y sirve como referencia o incluso se puede incorporar tal cual en el software que los deba implementar.

En ambos casos, las implementaciones FLOSS sirven de validación y también, al ser un código disponible y modificable, sirven para adaptar un estándar a ciertas interpretaciones o traducirlos a otro estándar, actuando como el pegamento que junta dos sistemas normalmente incompatibles.

3. Herramienta comercial para clientes

Incluso para quienes no utilizan FLOSS, éste sigue teniendo alta relevancia como una herramienta comercial en dos aspectos:

1) Define el conjunto mínimo aceptable. Si debo pagar por una solución y perder autonomía, debe haber un valor agregado que justifique el sobrecosto frente a una alternativa FLOSS (con sus propios costos, pero también sus propias ventajas).
2) Presenta una alternativa real y usable a la cual acudir cuando hayan desacuerdos comerciales con proveedores TI; ya no es fácil para un proveedor estar en una posición de "mi solución es la única alternativa existente".

En base a lo anterior, queda claro que el rol del FLOSS es relevante y debe haber un interés por mantenerlo vivo. Eso no necesariamente implica financiar los proyectos, sino simplemente evitar romper el equilibrio del ecosistema en el cual funcionan los proyectos FLOSS.

Es importante recordar que en el FLOSS, los temas de autonomía, cumplimiento de estándares e interoperatividad son inherentes y los intereses de los creadores del software están perfectamente alineados con los de los usuarios, porque esencialmente tienden a ser las mismas personas. En cambio, en el software comercial, los proveedores siempre tienen el incentivo perverso de intentar crear diversos lock-in contra los cuales los clientes deben luchar. Por lo mismo, aunque hoy en día esos incentivos no se traducen en malas prácticas comerciales, es bastante lógico pensar que sin la existencia del FLOSS el escenario actual sería diferente. Por ende, el FLOSS cumple un rol similar a la milicia en los tiempos de paz.

CAMBIOS Y DESAFÍOS

Hoy en día, sobre todo en los proyectos exitosos, una preferencia por FLOSS no se justifica por grandes diferencias en costos. Los grandes cambios y desafíos tienen más relación con el potencial cambio en la forma en la cual la TI llega a los usuarios. Tiene mucho sentido que nos acerquemos cada vez más a un cobro por servicio en lugar de cobros de licencias que se basan en el derecho de autor que en realidad regula la copia, y sólo de forma indirecta mediante construcciones legales tiene incidencia sobre el uso ("si no lo usas de la forma que yo digo, te quito el derecho de haber hecho y/o utilizar la copia").

Por lo mismo, todos los conceptos ligados a entregar el software como servicio, tales como el cloud computing, generan un cambio en la forma de uso y es ahí donde se concentran los principales cambios y desafíos para la industria TI y el FLOSS naturalmente no escapa a ello.

Posiblemente también aumenten a futuro las discusiones sobre si el derecho de autor es efectivamente la mejor forma de promover el desarrollo de las TI, desde siempre han habido defensores de que un modelo más parecido al patentamiento sería mejor.

Infraestructura base

En esto el FLOSS tiene bastante que ganar. Cuando se habla que el 50% del poder computacional en el mercado es comprado por Google, Microsoft y Yahoo!, para todas ellas salvo una, el costo de licenciamiento por copia de software se vuelve un tema que crece exponencialmente. Así que, contrario a lo que se podría pensar, que el FLOSS es para las PYME que no tienen cómo financiar el par de licencias de software comercial que podrían ocupar, el tema de uso de FLOSS es mucho más relevante para las empresas que deben replicar su solución en miles o incluso millones de computadores que ofrecen sus servicios, en paralelo, pero donde necesitan una licencia por cada uno de ellos. Lo mismo sucede cuando uno mira qué sistemas utilizan los supercomputadores listados en el "Top 500", que en la mayoría de los casos utilizan FLOSS.

Neutralidad en el tratamiento de datos

No me refiero a neutralidad tecnológica, que es en sí mismo un oxímoron, sino neutralidad respecto del tratamiento de información. Hoy en día por ejemplo asumimos que existe neutralidad en la información que manejan y nos entregan los buscadores, aunque no tenemos herramientas para validar que ello efectivamente ocurra. Posiblemente en esa área exista la posibilidad de pensar en alternativas que sigan la motivación de FLOSS pero en relación a los servicios, en conjunto con Open Data, posiblemente con limitaciones de escala al menos por ahora. Sin embargo, ya existen ejemplos

concretos de iniciativas que van en esa línea: Wikipedia, OpenStreetMap y algunos proyectos de redes sociales.

FLOSS entregado como servicio

El caso de servicios que se basan en FLOSS, en particular FLOSS que utiliza alguna licencia que incorpora la retribución, así como la GPL, se genera un fenómeno interesante. En estricto rigor, el objetivo de la reciprocidad es evitar que alguien pueda tomar un software que es esfuerzo de una comunidad, hacer algunas modificaciones y luego adueñarse y/o lucrar con el resultado (una obra derivada en términos legales) sin entregarle esas modificaciones a nadie. Para no exagerar las limitaciones, y en parte porque también hay ciertas limitaciones prácticas, se dispuso que la reciprocidad entre en funcionamiento al momento de entregar el software a un tercero para que lo use. De esa forma no era necesario entregar modificaciones privadas, pero se resguardaba el acceso al código fuente para toda persona que usara el software.

Pero cuando para usar el software no es necesario tener una copia, sino por ejemplo acceder al software vía Web, cambia el modelo y ahora es posible tomar un software GPL, hacerle modificaciones y entregar un servicio mediante ese software modificado sin ninguna restricción de entregar esas modificaciones al menos a quienes utilizan el software.

Como respuesta legal a ese problema surgen licencias alternativas como la Affero, que hace más severa la cláusula de reciprocidad de la GPL al hacerla mandatoria al momento de ofrecer un servicio basado en un software bajo licencia Affero. De esta forma, un usuario de un software vía Web u otro mecanismo similar tiene el derecho de obtener una copia del código fuente, montar su propio servicio y/o verificar el funcionamiento del software.

En este ámbito aún no se llega a un mecanismo maduro y estable para manejar tanto la gestión del software mismo como la disponibilidad de los datos.

> Hoy en día, sobre todo en los proyectos exitosos, una preferencia por FLOSS no se justifica por grandes diferencias en costos. Los grandes cambios y desafíos tienen más relación con el potencial cambio en la forma en la cual la TI llega a los usuarios.

Eficiencia energética

Este es uno de esos temas de fácil solución técnica, pero que requiere una decisión muchas veces corporativa que puede ser más esquiva de lo razonable.

Contrario a lo que se podría pensar, la industria TI no es tan limpia como parece. Sin considerar la huella de carbono de la fabricación de los computadores sino tan sólo el consumo eléctrico, la huella de carbono de la operación de las TI hoy en día según varios especialistas supera la de la industria de aviación. Más aún en Chile, donde la energía eléctrica hoy es sinónimo de combustibles fósiles para su generación.

La eficiencia energética de un sistema requiere una alta coordinación entre el software y el hardware. A nivel macro (sistema operativo) no hay demasiadas diferencias, pero sí pueden haber diferencias muy notorias en el consumo que tiene un dispositivo, por ejemplo, una tarjeta de red o una de vídeo, cuando es utilizado mediante un driver privativo (o propietario si se prefiere llamar así) o mediante un driver Open Source creado con la poca información pública y obtenido mediante ingeniería reversa.

Independiente que lo consideremos justificado o no, el hecho objetivo es que hay mucha reticencia, en particular en la industria de tarjetas de vídeo, a entregar información detallada del funcionamiento del hardware. Esto puede llevar a que en algunos casos una diferencia entre un driver y otro pueda tener consecuencias más indirectas que las obvias de correcto funcionamiento y usabilidad.

CONCLUSIONES

El tema del FLOSS sigue tan vigente como antes, y tal como los temas relevantes y los desafíos de la industria TI van fluctuando (según muchos se van repitiendo pero con diferentes nombres y quizás diferentes énfasis), así también ocurre con el FLOSS. Hoy en día hay menos discusión absolutista, y más consideración racional y objetiva, lo cual a mi juicio es sano.

Asimismo, existe participación por parte de empresas de todo tipo y tamaño como parte de los ecosistemas que mantienen funcionando a los proyectos FLOSS, y con ello aumentan las herramientas disponibles para enfrentar los problemas y desafíos que tendrán estos proyectos a futuro. Incluso pienso que a futuro debiéramos ver en los concursos por fondos públicos, que disponer del software y/o datos generados durante la ejecución bajo parámetros de apertura (FLOSS, Open Data) tenga tanto o más connotación positiva como la actualmente exigida "protección de propiedad intelectual" de los resultados, que en la práctica significa restringir al beneficiario de los fondos la explotación comercial futura de los resultados. BITS

OpenStreetMap: **el mapa libre del mundo**

Julio Costa
Administrador de Negocios Internacionales, Universidad de Valparaíso, Chile. Asesor en e-learning, Academia Nacional de Estudios Políticos y Estratégicos, Ministerio de Defensa Nacional, Chile. Fundador de zambelliknowledge. Presidente del Directorio, OpenStreetMap Chile; becario OSI+OSMF, State of The Map 2009, Amsterdam.
julio.costa@openstreetmap.cl

OpenStreetMap es a la vez un proyecto de cartografía libre y una comunidad de voluntarios que funciona con una mecánica colaborativa. La calidad de sus datos es supervisada por los mismos contribuyentes.

Además de esta mecánica, comparte con otros proyectos de tipo "wiki", el escepticismo inicial de quienes no han visto cómo, en la práctica, el vandalismo resulta marginal y tiende a volverse inocuo frente al avance de las herramientas y estructuras destinadas a contrarrestarlo, pero principalmente frente al fortalecimiento de comunidades comprometidas con el cuidado y mejora de un sector del mapa.

En 2004, cuando Wikipedia recién pasaba el millón de artículos –quince veces la cantidad que Encarta, el barco insignia de las enciclopedias electrónicas privativas de los años noventa, llegó a tener antes de su cierre definitivo- Steve Coast, en aquel entonces un estudiante de Ciencias de la Computación del University College of London, decidía que las barreras de entrada que la cartografía comercial imponía a los emprendedores y creadores eran demasiado altas y se debía hacer algo para cambiar eso. Es así como desarrolló la idea y comenzó la construcción de las herramientas necesarias para crear una base de datos cartográfica libre, paralela a la del Ordnance Survey, institución gubernamental que por 220 años había prácticamente monopolizado el desarrollo de la gran cartografía en el Reino Unido.

Si bien el nacimiento del proyecto se explica por múltiples causas, destacan el modelo de negocios que el Ordnance Survey utilizó durante décadas y que, empezando el siglo XXI, resultaba una carga demasiado pesada para los nuevos emprendedores, programadores, diseñadores, y autoridades. Éste se concentraba en una estrategia de altos

precios y una férrea política de aplicación de sus "Derechos de Autor".

Además de estas barreras legales y financieras, la cartografía del Ordnance Survey también adolece de un defecto común en la inmensa mayoría de los mapas actuales: la capacidad y velocidad en la corrección de errores está supeditada a los ciclos de edición de nuevos mapas.

Fueron estas características -altos precios, licencias sumamente restrictivas, sumados a la rigidez y lentitud de los cambios- los que llevaron a Coast a presentar su nueva idea en EuroFOO un 20 de agosto de 2004. Sólo algunas semanas antes, el 9 de agosto, había registrado el dominio openstreetmap.org, y un par de semanas después, a principios de septiembre, se publicaría el primer mensaje en la lista de correo.

OBJETIVOS

OpenStreetMap es un proyecto destinado a generar y poner a disposición del público datos geográficos libres.

Aunque su gratuidad y licenciamiento (actualmente bajo Creative Commons BY-SA, en proceso de transición a Open Database License-ODbL) son características destacables, el proyecto no se trata únicamente de eso. Una de las ideas fundamentales es dar acceso total a los datos "subyacentes", de manera que cualquier persona pueda hacer desarrollos innovadores y creativos sin tener como límite el contar con acceso sólo a datos preprocesados. Esto, que para algunos puede sonar como una barrera de entrada para desarrolladores menos avezados, es en realidad una ventaja fundamental, que permite a los creadores no sólo controlar el estilo y tipo de renderizado que se hace, sino también los criterios que usaremos a la hora de generar rutas para distintos medios de transporte, modelos en 3D, cartografía para Dispositivos de Navegación Personal, o el Tesauro de términos a los cuales responderá un motor de Geocoding.

Entonces, entendemos por Libre el acceso total y permanente a los datos y no simplemente la gratuidad parcial que

CC-BY Andrew Turner.

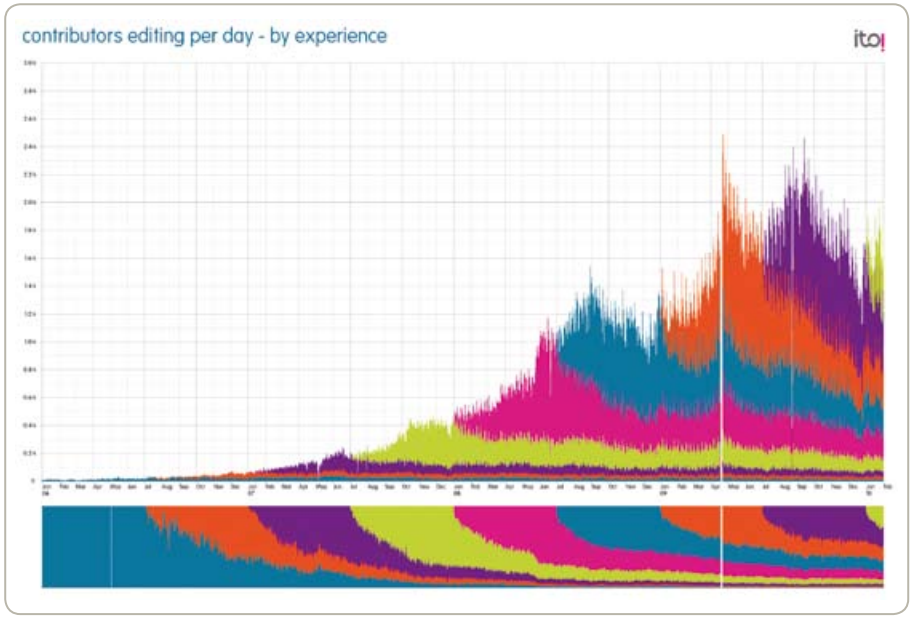
ofrecen "alternativas" como Google Map Maker, Waze o Wikimapia.

No se está trabajando gratis para alimentar una Base de Datos corporativa, algo por cierto legítimo en la medida que se entienda a cabalidad lo que se está haciendo, sino que se está contribuyendo a poblar y mejorar un repositorio de conocimiento universalmente disponible.

Este acceso total se logra normalmente a través de un archivo generado semanalmente llamado planet.osm, que contiene la totalidad de la Base de Datos de OpenStreetMap y del cual se pueden sacar extractos parciales, sumado a archivos XML diferenciales relativamente pequeños, que son publicados en ciclos de un minuto, una hora y un día. Estos archivos "diff" permiten mantener una base de datos actualizada sin necesidad de recargar la totalidad del planet.osm y la correspondiente espera de una semana.

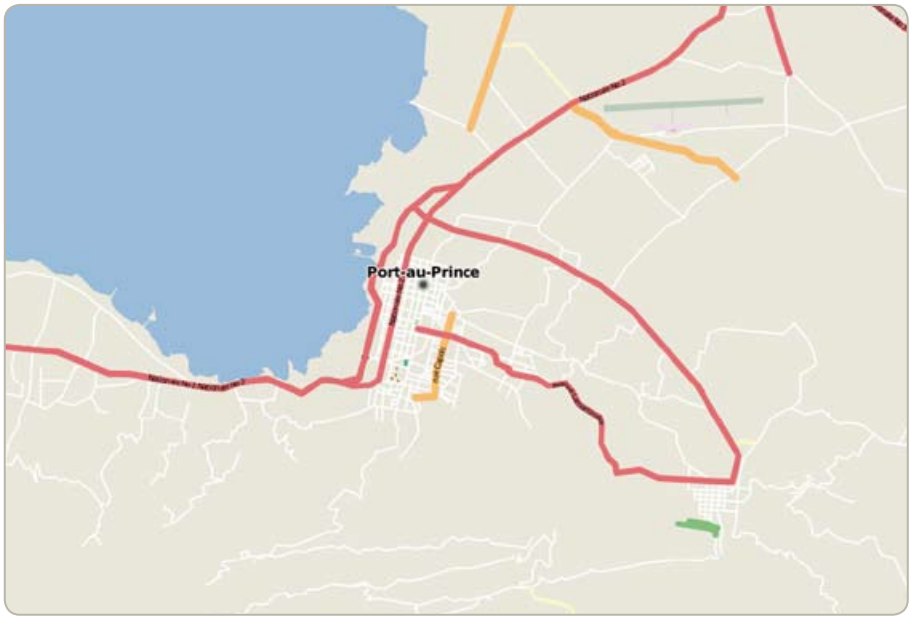
También se puede acceder a los datos a través de la API de OpenStreetMap, pero ésta no trabaja sobre grandes áreas, pues está pensada para actuar como "intermediario" de lectura y escritura en las operaciones regulares de edición.

Figura 1



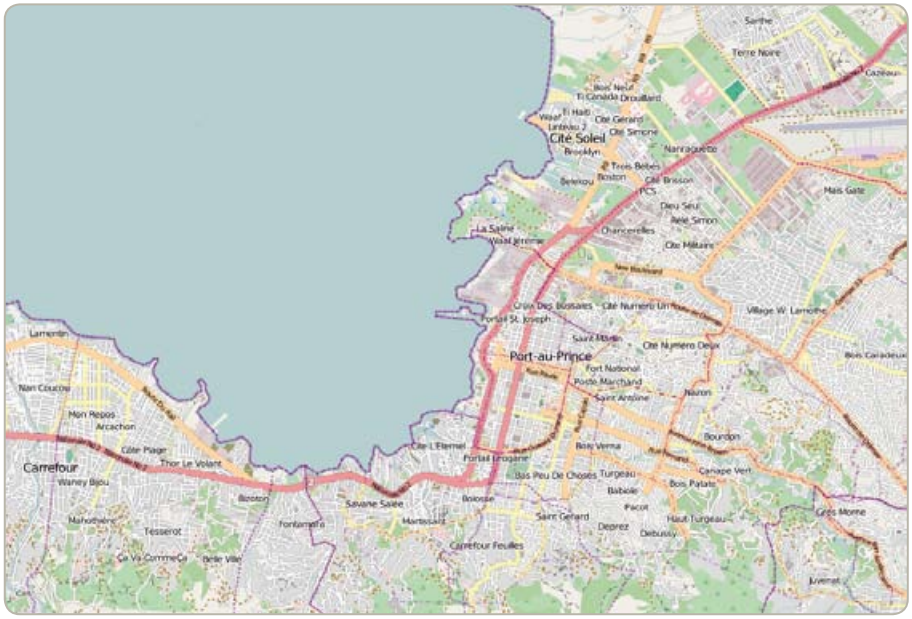
CC-BY-SA www.itoworld.com

Figura 2



CC-BY-SA www.itoworld.com

Figura 3



Al trabajar con receptores GPS de uso civil se tiende a pensar que la precisión no será "razonable", pero lo cierto es que un nivel de precisión de digamos cinco metros, es mucho más útil de lo que cualquiera podría pensar. Quizás no para ubicar un misil Tomahawk en un punto preciso de un búnker, pero sí para georreferenciar direcciones, representar usos de suelo, mostrar el transporte público en movimiento por las calles, para uso en Dispositivos de Navegación Personal, y de muchas otras formas. La recomendación general para los contribuyentes más concienzudos es trazar varias veces una ruta, idealmente en distintos horarios y días, lo que expone tu receptor GPS a distintos satélites y condiciones ambientales, consiguiendo buenos "promedios", los que se traducen en vías más precisas.

COMUNIDAD

Los datos del mapa tienen poco o nulo valor si no existiera la comunidad que está detrás de éste. Son ellos quienes mantienen el mapa en constante evolución, quienes "mapean" hasta los más extraños detalles, y quienes alertan y revierten el poco vandalismo que efectivamente se observa.

En este momento el número de contribuyentes inscritos se empina por sobre las 440.000 personas, de las cuales entre un 5% y 10% colabora de forma constante. En la Figura 1 se puede apreciar el crecimiento en la cantidad de ediciones diarias entre enero de 2006 y enero de 2010, y la contribución en términos de grupos de "experiencia", segmentados por semestre.

Es natural que cada grupo/semestre nuevo aporte proporcionalmente más, pues existe un crecimiento exponencial del universo de usuarios.

Las herramientas de primera línea para comunicarse y aprender sobre OpenStreetMap son las listas de correo y el wiki, pero también existen eventos regulares como los Mapping Parties y la conferencia anual de OpenStreetMap, el State of The Map, que celebró su quinta versión este año en Denver.

La OpenStreetMap Foundation fue creada en agosto de 2006 con el objetivo de: custodiar el dominio, los servidores, y servicios necesarios para que OpenStreetMap funcione; ofrecer un cierto nivel de "aislación jurídica"; y proveer un "vehículo" para la recolección de fondos para el proyecto. Teniendo como misión genérica el apoyo al desarrollo de datos geográficos libres.

PRO BONO PUBLICO

Toda la comunidad -desde los miembros del directorio de la OpenStreetMap Foundation hasta los contribuyentes individuales, pasando por los Sysadmin que sostienen la

> Una de las ideas fundamentales es dar acceso total a los datos "subyacentes", de manera que cualquier persona pueda hacer desarrollos innovadores y creativos sin tener como límite el contar con acceso sólo a datos preprocesados.

Figura 4

CC-BY-SA www.itoworld.com

infraestructura del proyecto, y los miembros del Humanitarian OpenStreetMap Team-colabora de forma voluntaria. No se necesita ser experto en Rails o en Organizaciones No Gubernamentales y Cooperación Internacional para participar, basta con tener buenas intenciones y tiempo para contribuir.

El proyecto nació con una orientación implícita a facilitar el trabajo de desarrolladores y emprendedores, pero se ha ido diversificado hasta un punto tal en que personas de todo el mundo construyen proyectos y soluciones humanitarias basadas en el mapa. Sólo algunos ejemplos de esto son el Humanitarian OpenStreetMap Team (HOT), Map Kibera, Afghanistan Election Data, y la utilización como capa por defecto en implementaciones de Ushahidi para Haiti, Japón y Libia, entre otros.

Uno de los fenómenos humanitarios más destacables es el que se dio luego del Terremoto de enero de 2010 en Haití. Grupos de contribuyentes se organizaron para mapear de manera remota las áreas más afectadas. Se contactó a los grandes proveedores de imágenes satelitales (GeoEye y DigitalGlobe), los cuales liberaron imágenes de alta resolución tan pronto como sus satélites estuvieron en posición. Se generaron teselas a partir de esas imágenes, las que estuvieron disponibles para ser usadas como fondo en Potlatch y otros editores, por usuarios de todo el mundo. Esto llevó a que un mapa relativamente pobre (Figura 2), se convirtiera en el mapa más completo de Haití (Figura 3) en cuestión de días.

Grupos de rescatistas, así como el sitio de Ushahidi para Haití, pudieron usar estos mapas sin problemas de licenciamiento, con constantes actualizaciones, y con la rápida implementación de características especiales necesarias para ese tipo de emergencia en particular, como la geolocalización de campamentos "espontáneos" (Figura 4).

DESAFÍOS TÉCNICOS

Si bien los datos "crudos", que son el centro del proyecto, se pueden adaptar para funcionar bajo múltiples escenarios, existen configuraciones comunes de trabajo.

Por ejemplo, un servidor para el renderizado de teselas (tiles), normalmente involucra una distribución de Linux para servidores, con PostgreSQL/PostGIS como motor de Base de Datos, junto con Python y algunas otras librerías necesarias para hacer funcionar Mapnik, el software libre de renderizado más utilizado para convertir los datos geográficos, almacenados en Bases de Datos o Shapefiles, a bitmaps.

Mapnik es una evolución de los motores de renderizado libres y la diferencia de éste con motores arcaicos como Osmarender, aún puede apreciarse al seleccionar esta segunda alternativa en el menú de capas de la página principal de OpenStreetMap.

La Base de Datos PostgreSQL es poblada a partir de un archivo planet y los diferenciales ya mencionados, o un extracto del primero, utilizando una aplicación conocida como *osm2pgsql*.

Es común que las teselas sólo sean renderizadas previamente en los niveles menos profundos de zoom (0-11), dejando los niveles más profundos (12-18), que involucran muchísimas más imágenes para la misma superficie (el número aumenta por un factor de cuatro por cada nivel de zoom que avanzamos. Ver tabla a continuación), para ser renderizadas bajo demanda, utilizando *mod_tile*, un módulo de Apache, para gestionar el caché, la expiración de los archivos, y los requerimientos de nuevas imágenes al "back-end" (normalmente renderd o tirex).

> Si bien a nivel nacional se han hecho algunas importaciones masivas de datos, el grueso del trabajo se sigue haciendo de forma manual, aportando de esa manera a una revisión más minuciosa de los datos que se almacenan y actualizan.

Si todas las potenciales teselas se renderizaran previamente, y tomando en consideración que en promedio tienen un tamaño de 633 bytes, se necesitarían un poco más de 54.000 Gigabytes para almacenarlas. La realidad es que con el esquema de almacenaje y renderizado utilizado, y a pesar de la alta demanda que tienen los servidores de la OpenStreetMap Foundation, sólo se utiliza un poco menos de 1.000 Gigabytes, un 1,79% del máximo potencial.

Pero no debemos quedarnos sólo con este dato. Cada hora los servidores de la fundación entregan cerca de cuatro millones de teselas, llevando este número a más de mil millones en cerca de once días. Algo que a la iniciativa "libre" del Ordnance Survey, OS Openspace, le toma cuatro días y más de tres años respectivamente.

Y todo este trabajo se sigue realizando sobre Hardware relativamente simple y financiado a través de donaciones. Por ejemplo *yevaud,* el servidor que se encarga del renderizado y entrega de tiles utilizando Mapnik, cuenta con dos procesadores Xeon de cuatro núcleos, 48GB de RAM, y múltiples arreglos de discos SATA.

Zoom	Teselas renderizadas	Maximo (4^zoom)	% Renderizado/ Maximo
0	1	1	100
1	4	4	100
2	16	16	100
3	64	64	100
4	256	256	100
5	1,024	1,024	100
6	4,096	4,096	100
7	16,384	16,384	100
8	65,536	65,536	100
9	262,144	262,144	100
10	1,048,576	1,048,576	100
11	4,194,304	4,194,304	100
12	13,475,072	16,777,216	80.32
13	35,640,512	67,108,864	53.11
14	87,820,928	268,435,456	32.72
15	163,872,384	1,073,741,824	15.26
16	287,448,064	4,294,967,296	6.69
17	429,535,936	17,179,869,184	2.50
18	617,515,264	68,719,476,736	0.90
Total	1,640,900,565	91,625,968,981	1.79

EN CHILE

La comunidad local se ha desarrollado paulatinamente durante los últimos cuatro años, con algunos extranjeros residentes en Chile inicialmente, pero con una creciente e intensiva participación de chilenos actualmente.

Se han organizado múltiples eventos pequeños del tipo Mapping Party, pero también eventos de mayor convocatoria como la "Mañana de Mapas Libres". Además se participa de manera constante en conferencias y ferias anuales como el Día del Software Libre, la FLISOL y el Encuentro Linux, entre otros.

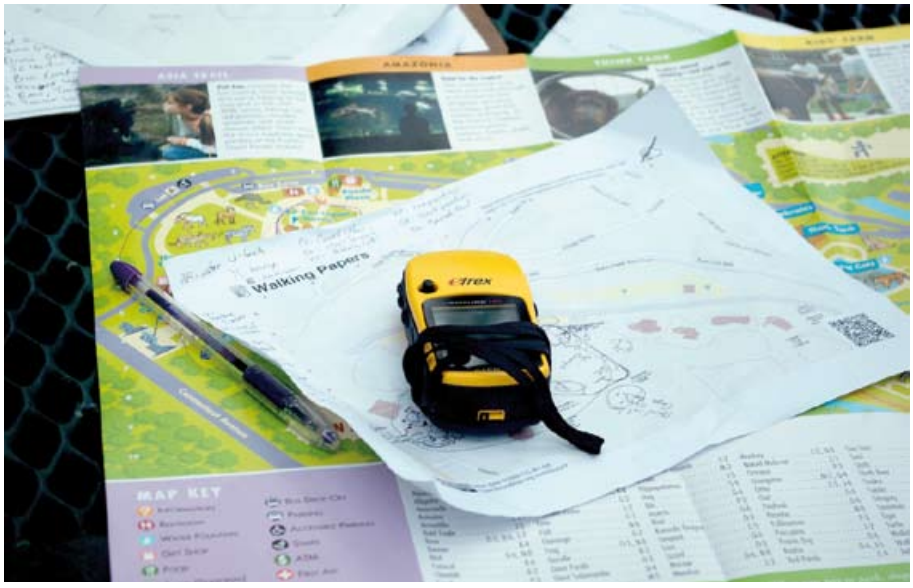
Si bien a nivel nacional se han hecho algunas importaciones masivas de datos, gracias a la colaboración de instituciones como la Dirección de Vialidad del Ministerio de Obras Públicas, el Gobierno Regional Metropolitano de Santiago, el Instituto Nacional de Estadística (INE), la Secretaría de Planificación de Transporte (SECTRA), la Coordinación de Transportes de Santiago, entre otros, el grueso del trabajo se sigue haciendo de forma manual, aportando de esa manera a una revisión más minuciosa de los datos que se almacenan y actualizan.

A diferencia del Ordnance Survey en su momento, el Instituto Geográfico Militar de Chile, ha tomado contacto con OpenStreetMap en momentos en que el concepto de OpenData está mucho más difundido y es más apreciado, incentivando la colaboración y anunciando la liberación de varios conjuntos de datos en el futuro cercano.

CÓMO EMPEZAR

Normalmente comenzamos creando una cuenta de usuario en el sitio principal de OpenStreetMap y familiarizándonos con el editor más simple, Potlatch. Este se encuentra disponible a través de la pestaña "Editar" en ese sitio, y sólo requiere tener instalado el plugin de Adobe Flash en nuestro navegador. A través de él se puede editar sobre imágenes satelitales y aéreas

CC-BY-SA www.itoworld.com.



CC-BY Andrew Turner.

liberadas por distintas instituciones, así como múltiples renders que pueden servir de referencia. Cuenta con un extenso listado de elementos cartográficos predefinidos: tipos de rutas, hidrología, usos de suelo, puntos de interés, etc.

Una vez que estamos familiarizados con alguno de los editores (Potlatch, JOSM, Merkaartor, Mapzen, etc.), podemos salir a terreno con nuestro GPS y generar un track en formato GPX, el cual luego subiremos al sitio de OpenStreetMap, en la pestaña "Trazas GPS".

Una vez que el trazo está cargado en los servidores de OSM podemos editar sobre él, lo que en la práctica nos permite llegar a lugares en los que las imágenes satelitales y aéreas muchas veces no son lo suficientemente buenas, por ejemplo, senderos de parques que son tapados por los árboles, o calles que son tapadas por sus propios edificios en las fotografías (depende del "azimuth" de esa fuente).

También existe la opción de usar Walking-Papers, una herramienta desarrollada por Mike Migurski de Stamen Design en San Francisco, pensando en aquellos lugares donde la gente no puede pagar por un receptor GPS y los programas GPStogo de préstamo de equipos simplemente no dan abasto.

Sólo se debe ingresar a http://walking-papers.org/, seleccionar un sector del mapa con el mayor zoom posible (en mi experiencia, menos de 16 o 17 no resulta útil), se selecciona una orientación, una distribución, y un proveedor para el mapa, y se hace clic en "Crear". Esto generará un archivo PDF para imprimir, el que tendrá marcadores en sus cuatro esquinas, dos de los cuales serán un Código QR y el logotipo de la licencia Creative Commons. Estos permiten que una vez que hicimos todo tipo de anotaciones sobre el mapa, escaneamos éste, y lo subimos al sitio de Walking Papers, el sistema pueda identificar su ubicación y escala, y lo use como fondo en un Potlatch en el mismo sitio.

Si quieres practicar el uso del editor, no hay problema, sólo recuerda no guardar al terminar de hacerlo. Los datos son transferidos a la API y a la Base de Datos, sólo cuando presionas "Guardar", comentas el Changeset y aceptas.

Ante preguntas sobre cómo etiquetar nodos, vías y áreas, o sobre cómo utilizar los editores u otras tecnologías, no dudes en preguntar en las listas de correo o revisar el wiki, la fuente "inagotable" de conocimientos y repositorio de la mayor parte de los temas en que hay convenciones sobre cómo hacer las cosas. BITS

REFERENCIAS

http://www.openstreetmap.cl/

http://www.openstreetmap.org/

http://wiki.oreillynet.com/eurofoo/index.cgi

http://lists.openstreetmap.org/listinfo

http://wiki.openstreetmap.org/wiki/Main_Page

http://2010.afghanistanelectiondata.org/

http://www.osmfoundation.org

http://mapnik.org/

http://mike.teczno.com/notes/walking-papers.html

Análisis de Datos Astronómicos

Karim Pichara
Profesor Asistente, DCC Pontificia Universidad Católica de Chile. Investigador del Centro de Astro-Ingeniería y del grupo de Biomedicina de la Pontificia Universidad Católica de Chile. Doctor en Ciencias de la Ingeniería, Pontificia Universidad Católica de Chile (2010). Posdoctorante en el laboratorio "Time Series Center" del Centro de Astrofísica de la Universidad de Harvard (2011-2013).
kpb@ing.puc.cl

Rodolfo Angeloni
Investigador posdoctoral, DAA Pontificia Universidad Católica de Chile. Doctor en Astronomía (2009), Universitá di Padova, Italia . Investigador Responsable del Proyecto FONDECYT N. 3100029 "Topics in Stellar Variability: from VISTA to ALMA".
rangelon@astro.puc.cl

Susana Eyheramendy
Profesora asistente, Depto. de Estadística Facultad de Matemáticas, Pontificia Universidad Católica de Chile. PhD Depto. de Estadística Universidad de Rutgers, EE.UU; posdoc Universidad de Oxford y Ludwig-Maximilian Universitat/Instituto de Epidemiología del Helmholtz Zentrum Munich, Alemania. Su investigación se basa en el análisis y desarrollo de métodos en estudios genéticos de asociación y en aplicaciones de métodos de minería de datos a problemas astronómicos.
susana@mat.puc.cl

En el Centro de Astro-Ingeniería de la Pontificia Universidad Católica de Chile, un grupo de científicos conformado por los profesores Márcio Catelan, Andrés Jordán y Rodolfo Angeloni de Astronomía; Susana Eyheramendy de Estadística; Karim Pichara de Ingeniería en Ciencia de la Computación, y el alumno de Ingeniería Cristóbal Berger, se dedican al desarrollo de herramientas inteligentes para el análisis de Datos Astronómicos. Durante los últimos años, ha existido un creciente interés en aplicaciones de inteligencia artificial (Russel and Norvig (2010)) y aprendizaje de máquina (Mitchel (1997)) para la investigación astronómica debido al gran desarrollo tecnológico de los telescopios, cada vez capaces de generar una mayor cantidad de información imposible de ser analizada en su totalidad por humanos. Por ejemplo, el próximo telescopio LSST ("Large Synoptic Survey Telescope"[1]) tendrá la labor de producir durante diez años alrededor de 30 Terabytes diarios de información proveniente del Universo, esto corresponde a varios billones de objetos, cada uno observado en alrededor de 1.000 instantes distintos de tiempo. El proyecto "Vista Variables in the Via Lactea (VVV) ESO Public Survey"[2], escaneará la Vía Láctea arrojando mediciones en la banda infrarroja de más de diez mil millones de objetos en el espacio. Estos desarrollos impulsan nuevas necesidades científicas: a mayor cantidad de información disponible, mayor es la necesidad de nuevas tecnologías para el análisis de estos datos.

Una de las tareas más importantes en el análisis de datos del espacio es la clasificación automática de objetos estelares. Existe hoy gran interés en desarrollar modelos de

1 http://www.lsst.org/lsst/
2 http://mwm.astro.puc.cl/mw/index.php/Main_Page

Figura 1

Segundo lanzamiento de la imagen del VST, probablemente es el mejor retrato del cúmulo globular Omega Centauri que alguna vez se haya obtenido. Omega Centauri, en la constelación de Centaurus es el cúmulo globular más grande del cielo.

Cuando sea observado con el telescopio infrarrojo VISTA, nuestro grupo logrará obtener curvas de luz para un número importante de diferentes tipos de estrellas variables. Estas observaciones constituirán una fracción importante del conjunto de entrenamiento que estamos construyendo en el proyecto "VVV Templates".

aprendizaje de máquina capaces de aprender a clasificar automáticamente estos objetos a partir de bases de datos previamente rotuladas por astrónomos (Debosscher et al. (2007), Dubath et al. (2011), Richards et al. (2011), Kim et al. (2011)). Estos sistemas de clasificación deben considerar desde el preprocesamiento de los datos hasta la generación del modelo capaz de clasificar automáticamente los objetos.

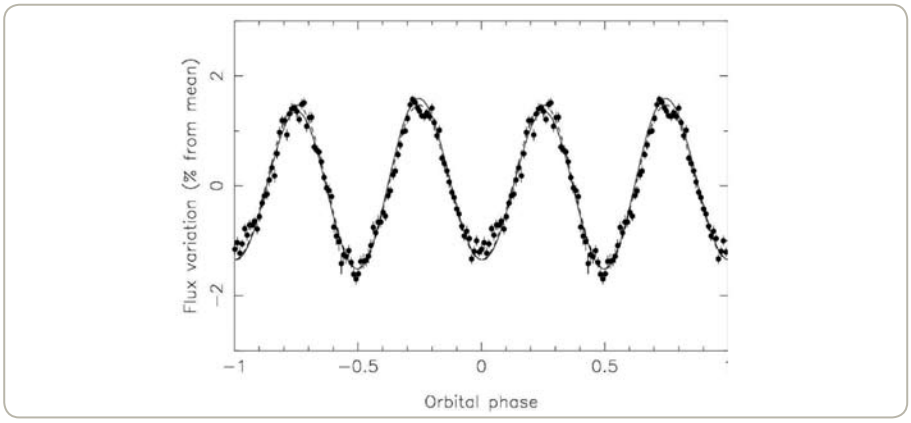
Dado que la mayoría de los proyectos observacionales como LSST y VVV incluyen observar estrellas variables (estrellas que muestran una variación en su brillo en función del tiempo. Ver catálogo de las distintas clases en http://www.sai.msu.su/gcvs/gcvs/iii/vartype.txt), es natural enfocar los esfuerzos en el análisis de series de tiempo o curvas de luz (gráfico que se obtiene de la variación del brillo en función del tiempo, Figuras 2, 3 y 5), este análisis busca representar en forma compacta una curva de luz de tal manera de simplificar la información que recibe un algoritmo de clasificación.

En la literatura existe una amplia gama de modelos de análisis de series de tiempo (Percival et al. (2003), Mills et al. (1990), Bloomfield (1976), Hamilton (1994)). La principal ventaja de usar modelos para analizar las series de tiempo es poder extraer características propias de la forma de cada curva de luz, de tal modo de obtener información útil para que los algoritmos de clasificación automática puedan desempeñar su labor usando como principal información estas características obtenidas del análisis de cada serie de tiempo. Existen numerosas técnicas para modelar curvas de luz (Lomb (1976), Scargle (1982), Ponman (1981), Kurtz (1985)). Estos modelos estiman los parámetros y frecuencias de un modelo armónico sobre la forma de la curva de luz, de tal modo de usar los parámetros encontrados como descriptores de cada curva.

Consideremos como $y(t)$ la intensidad de luz observada en un instante t, sea $\hat{y}(t)=a+bt$ una estimación lineal de $y(t)$ y sea $r(t)=y(t)-\hat{y}(t)$. Iteramos entre los siguientes pasos:

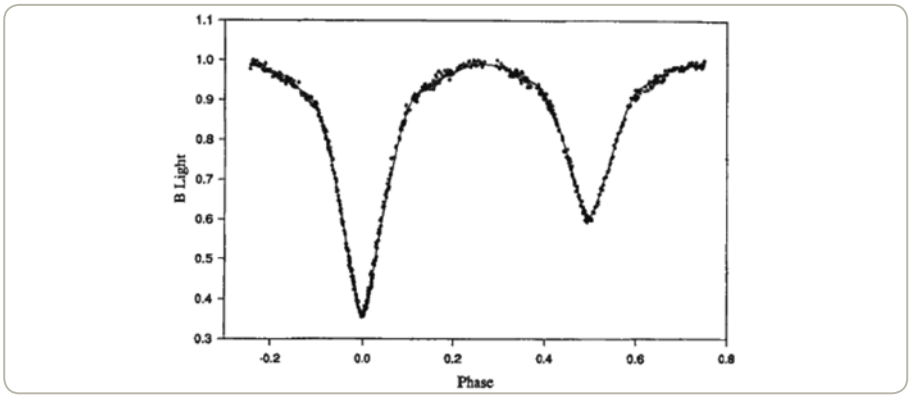
1. Realizar un análisis de Fourier para $r(t)$ con el objetivo de determinar cualquier periodicidad que podría existir usando el método Lomb-Scargle(Lomb (1976), Scargle (1982)). Una vez calculado el periodograma de Lomb-Scargle se selecciona el valor máximo. La frecuencia correspondiente

Figura 2



Curva de luz de KPD1930+2752 después de remover la señal debido a las pulsaciones de Billères et al. (2000) con un modelo de curva de luz (línea sólida) para la variabilidad elipsoidal asumiendo una inclinación de 90°.

Figura 3



Curva de luz de TT Aurigae observada por Wachmann, Popper, y Clausen (1986) y modelada por Terrell (1991).

Una de las tareas más importantes en el análisis de datos del espacio es la clasificación automática de objetos estelares. Existe hoy gran interés en desarrollar modelos de aprendizaje de máquina capaces de aprender a clasificar automáticamente estos objetos a partir de bases de datos previamente rotuladas por astrónomos.

f se usa para encontrar los parámetros de la siguiente función armónica, usando un método de mínimos cuadrados:

$$\hat{z}(t) = \sum_{j=1}^{m} (a_j \sin 2\pi f jt + b_j \cos 2\pi f jt) + b_0$$

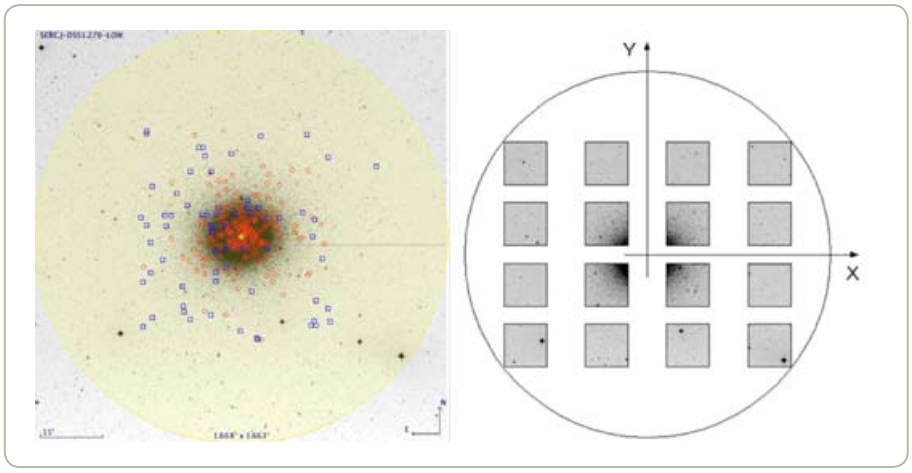
2. Actualizar $r(t)=r(t)-\hat{z}(t)$

En palabras, primero se resta la tendencia lineal de la serie de tiempo fotométrica, luego usando el periodograma de Lomb-Scargle identificamos el peak más alto y usamos la frecuencia correspondiente para ajustar un modelo armónico con m componentes. Esta nueva curva, junto con la estimación lineal son restadas desde la serie de tiempo y se busca una nueva frecuencia en el residuo usando el periodograma de Lomb-Scargle, la nueva frecuencia se usa para ajustar nuevamente el modelo armónico. Este proceso continúa hasta que se encuentran n frecuencias y se estiman n modelos armónicos con m componentes. Finalmente, las n frecuencias se usan para realizar el mejor ajuste a la curva de luz original:

$$\hat{y}(t) = \sum_{i=1}^{n} \sum_{j=1}^{m} (a_{ij} \sin 2\pi f_i jt + b_{ij} \cos 2\pi f_i jt) + a + bt$$

Las frecuencias f_i junto con los parámetros de Fourier a_{ij} y b_{ij}, constituyen el conjunto de parámetros con los cuales podemos representar las curvas de luz.

Figura 4



Próximas observaciones VISTA del cúmulo globular omega Cen. Panel izquierdo: un ejemplo de distribución a lo largo del cluster: los círculos rojos marcan la posición de estrellas RR Lyrae conocidas, los cuadrados azules marcan las posiciones de estrellas binarias eclipsantes, conocidas. Panel derecho: 16 detectores de VIRCAM@VISTA, con omega Cen en el centro del plano focal.

Uno de los principales problemas de esta representación es que los parámetros no son invariantes a traslaciones en el tiempo. En otras palabras, si de la misma estrella tenemos dos curvas de luz observadas para las cuales no coincide el instante de tiempo inicial, estas dos curvas de luz tendrán un conjunto distinto de parámetros representando la misma estrella. Para lidiar con este problema transformamos los coeficientes de Fourier en un conjunto de amplitudes A_{ij} y fases PH_{ij} como sigue:

$$A_{ij} = \sqrt{a_{ij}^2 + b_{ij}^2},$$

$$PH'_{ij} = \arctan(\sin(PH_{ij}), \cos(PH_{ij}))$$

donde:

$$PH_{ij} = \arctan(b_{ij}, a_{ij}) - \frac{jf_i}{f_1} \arctan(b_{11}, a_{11})$$

Notar que PH_{11} es elegido como la referencia y es igual a cero, además $PH`_{ij}$ toma valores en el intervalo $]-\pi,\pi]$

Una vez que tenemos una representación paramétrica para las curvas de luz, podemos utilizar modelos de clasificación para aprender a identificar cada tipo de estrella.

Existen numerosos algoritmos de clasificación automática en la literatura de Aprendizaje de Máquina (Mitchel (1997), Bishop (2006)). Los algoritmos de clasificación están basados principalmente en modelos matemáticos para encontrar espacios de separación entre las distintas clases de objetos, entre los algoritmos más nombrados están los "árboles de decisión" (Quinlan (1986)), "Support Vector Machines" (Cortes & Vapnik 1995), el clasificador "Naive Bayes" (Mitchel (1997)), el "clasificador de vecinos cercanos" (Mitchel (1997)), etc. El proceso de aprendizaje de estos algoritmos consta en utilizar un conjunto de instancias para entrenar (conjunto de entrenamiento) donde el algoritmo busca separar entre los elementos de distintas clases para luego probar el rendimiento del modelo de clasificación en un conjunto

de instancias que no fueron usadas en el proceso de entrenamiento (conjunto de evaluación). Es importante a la hora de iniciar el entrenamiento de un modelo de clasificación considerar que el objetivo final es que el modelo clasifique con un alto rendimiento las instancias del conjunto de evaluación, de tal modo de probar que el modelo es capaz de clasificar correctamente instancias nuevas, no procesadas durante el entrenamiento, eso asegura una buena capacidad de generalización del modelo de aprendizaje.

Al momento de lidiar con Datos Astronómicos aparecen muchas limitaciones que hacen más difícil el proceso de aprendizaje de clasificadores. Una de estas limitaciones corresponde al gran costo de obtener datos etiquetados para formar el conjunto de entrenamiento. Dado que los telescopios no arrojan información sobre el tipo de objeto que inspeccionan, sino que sólo información sobre algunas de sus características, es necesario que los astrónomos manualmente se dediquen a etiquetar estos datos de tal manera que un computador pueda iniciar el proceso de aprendizaje.

Dado también que algunos tipos de Datos Astronómicos existen hace muchos años, hoy gran parte de ellos se encuentran etiquetados y disponibles para la comunidad científica, éste es el caso de los datos ópticos. Lamentablemente existen otros tipos de datos que no están etiquetados por la comunidad astronómica, por ejemplo el VVV es la primera inspección de variabilidad estelar en la banda infrarroja, en este caso nuestro grupo debe lidiar con la construcción de una base de datos etiquetada de variabilidad estelar en la banda infrarroja. Así el principal propósito del proyecto "VVV Templates"[3] se traduce en construir un conjunto de entrenamiento en la banda infrarroja para los clasificadores automáticos, que hasta ahora sólo han sido utilizados en datos ópticos.

Parte del proyecto "VVV Templates" comprende observar el cúmulo globular Omega Cen (Figuras 1 y 4). Este cúmulo contiene varios millones de estrellas, entre ellas varios cientos de estrellas variables de diferentes tipos y nos permitirá obtener con una serie de observaciones una fracción importante de los datos que se necesitan para construir los "templates" de curvas de luz que se esperan obtener de este proyecto.

Con todos los recursos que se necesitan para obtener estos "templates" nace la necesidad de implementar modelos de clasificación con un nivel mayor de inteligencia, capaces de seleccionar eficientemente sólo los objetos más informativos de tal modo de aprender a clasificar con la menor cantidad de información posible, de tal modo de ahorrar recursos valiosos como el tiempo dedicado a las observaciones al telescopio. Este proceso de seleccionar instancias específicas es conocido en la literatura del aprendizaje de máquina como "active learning" o aprendizaje activo (Roy et al. (2001), Cebron et al. (2008), Baram et al. (2004)). Los modelos de aprendizaje activo van seleccionando en cada iteración la instancia más apropiada para el aprendizaje a partir de un conjunto de instancias candidatas, luego solicita a algún ente experto (en este caso el astrónomo) que clasifique la instancia previamente seleccionada, para luego incluir esta información en el conjunto de entrenamiento y refinar el clasificador y el modelo selector de instancias.

Más específicamente, considere un conjunto de instancias (curvas de luz) descritas por d parámetros $X=\{x_1,...,x_n\}$, donde $x_i \in R^d\, i \in [1,..,n]$ y un conjunto de C posibles clases de estrellas $Y=\{y_1,...,y_c\}$, donde cada x_i pertenece a una clase y_j, $j \in [1,..,C]$. Considere el conjunto $U \in X$ de curvas de

Figura 5

La "Nebulosa del Cangrejo del Sur", uno de los mejores ejemplos de estrella variable de tipo simbiótico en sus últimas fases de evolución. Esta imagen fue obtenida utilizando el telescopio espacial Hubble. Autores: Romano Corradi, Mario Livio, Ulisse Munari, Hugo Schwarz y NASA.

3 http://www.vvvtemplates.org/

luz no clasificadas y el conjunto $L \in XxY$ de curvas previamente etiquetadas (la clase de cada elemento en *L* es conocida).

El proceso de aprendizaje activo consiste en estratégicamente seleccionar curvas de luz desde el conjunto *U* para que sean etiquetadas por un astrónomo y luego agregadas al conjunto *L*. Cada vez que *L* cambia se actualiza un clasificador cuyo conjunto de entrenamiento es *L*. Principalmente la idea es seleccionar las curvas de luz que más aportan en el entrenamiento del clasificador. Sea *P(y|x)* la distribución (desconocida) que corresponde a la probabilidad de que un dato *x* pertenezca a la clase *y*, sea *P(x)* la distribución de probabilidades marginal sobre las instancias. Sea $\hat{P}_D(y|x)$ la distribución de probabilidades que el modelo clasificador debe aprender del conjunto de entrenamiento *D*. El error esperado de la clasificación es:

$$E_{\hat{P}_D} = \int_x Ls(P(x|y), \hat{P}_D(y|x))P(x)$$

Donde *Ls* es una función que mide el grado de pérdida o diferencia entre la distribución estimada y la distribución real:

$$Ls = \sum_y P(y|x)log(\hat{P}_D(y|x)))$$

El algoritmo de aprendizaje activo seleccionará la instancia x_i tal que al añadirla al conjunto *L* el clasificador entrenado con el conjunto resultante $L^* = L + (x_i, y_i)$ obtiene el menor error comparado con todas las otras instancias candidatas, es decir:

$$\forall(x,y)E_{\hat{P}_{L^*}} < E_{\hat{P}_L}$$

Lamentablemente la distribución real *P(y|x)* es desconocida, por lo tanto realizamos una estimación del error en base al valor esperado de la clasificación de cada instancia, usando el clasificador que se tiene hasta el momento evaluado sobre el conjunto de testeo, así el error estimado se calcula como:

$$\tilde{E}_{\hat{P}_{L^*}} = \frac{1}{|T|}\sum_x \sum_y \hat{P}_{L^*}(y|x)log(\hat{P}_{L^*}(y|x))$$

Donde $|T|$ es el número de elementos en el conjunto de testeo.

En palabras simples, el modelo va a elegir como siguiente instancia a la que más disminuye la incerteza del clasificador una vez agregada al conjunto de entrenamiento. Notar que la incerteza se mide como la entropía de la clasificación.

Hasta ahora se han obtenido resultados bastante positivos, la Figura 6 muestra la exactitud del clasificador a medida que van agregándose instancias con el proceso de aprendizaje activo. La línea azul corresponde a la precisión[1] y la línea roja al recall[2]. La línea recta bajo la línea azul corresponde a la

Figura 6

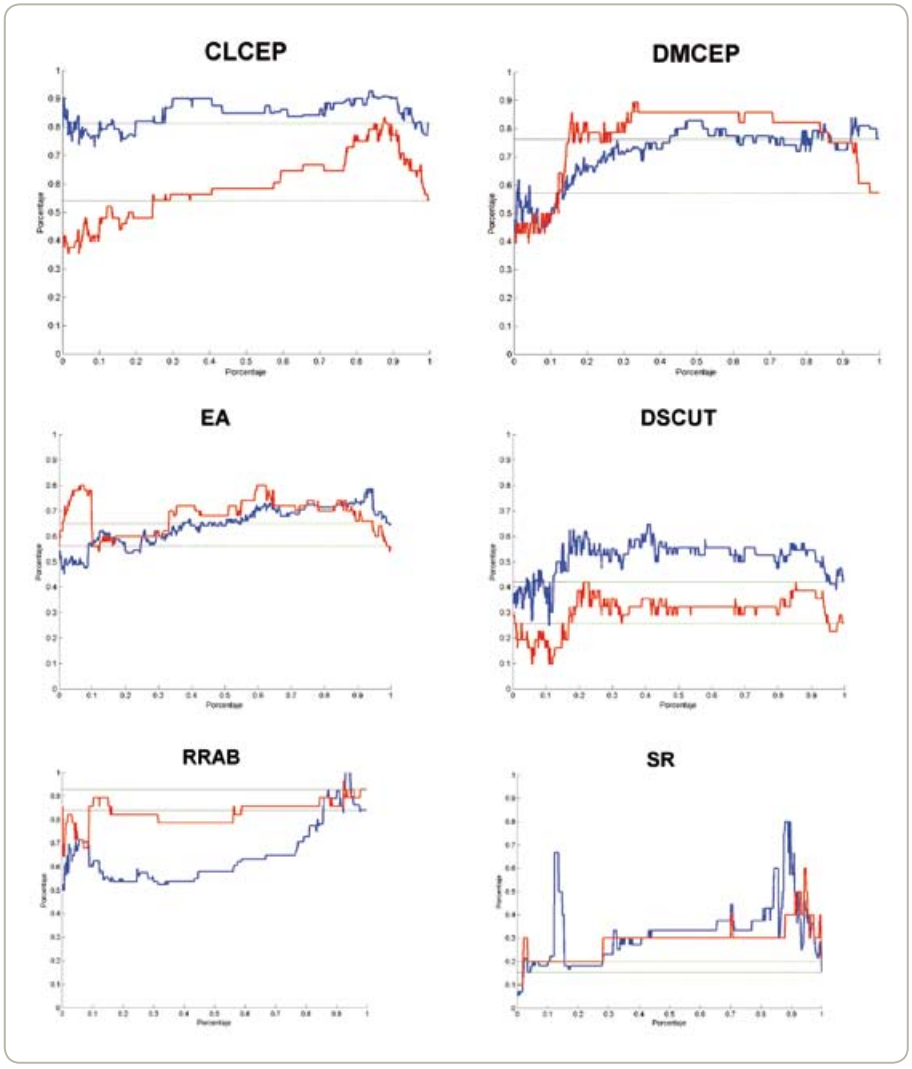


Gráfico que muestra la exactitud del clasificador a medida que van agregándose instancias con el proceso de aprendizaje activo. La línea azul corresponde a la precisión y la línea roja al recall. La línea recta bajo la línea azul corresponde a la precisión obtenida usando todo el conjunto de entrenamiento. La línea recta bajo la línea roja corresponde al recall obtenido usando todo el conjunto de entrenamiento. Cada vez que la línea roja (azul) está por sobre su línea recta implica que el recall (precisión) es mayor que el obtenido con el 100% del conjunto de entrenamiento.

1 Precisión: de los elementos que el clasificador dijo que eran de la clase en cuestión, cuántos realmente eran.
2 Recall: de los elementos de la clase que el clasificador tenía que identificar, cuántos identificó.

Susana Eyheramendy

Karim Pichara B.

Cristóbal Berger

Rodolfo Angeloni

Andrés Jordán

Márcio Catelan

precisión obtenida usando todo el conjunto de entrenamiento. La línea recta bajo la línea roja corresponde al recall obtenido usando todo el conjunto de entrenamiento. Cada vez que la línea roja (azul) está por sobre su línea recta implica que el recall (precisión) es mayor que el obtenido con el 100% del conjunto de entrenamiento. Se puede apreciar por ejemplo que en el gráfico de la clase CLCEP el clasificador logra igualar en recall y precisión los resultados obtenidos cuando se usó el 100% del conjunto de entrenamiento sólo usando un 30% de instancias, seleccionadas estratégicamente con el proceso de aprendizaje activo. Resultados similares se ven en las otras clases, excepto en la clase RRAB, donde el clasificador no puede igualar los resultados sino hasta llegar a seleccionar todas las instancias del conjunto de entrenamiento.

Los pasos siguientes de esta investigación comprenden desarrollar un modelo para automatizar la decisión de dónde detener el proceso de aprendizaje activo, es decir, en qué momento el clasificador debe decidir que ya aprendió lo suficiente y no necesita pedir la clasificación de más instancias. Para lograr este objetivo existen muchos desafíos por superar, entre ellos la inestabilidad de los resultados en algunos casos.

Existen muchas otras aristas de investigación que se irán explorando con el tiempo, este grupo científico pretende seguir creciendo y desarrollando nuevas tecnologías para el análisis de Datos Astronómicos, se espera en un futuro próximo poder contar con un número importante de estudiantes de posgrado realizando sus investigaciones en el Centro de Astro-Ingeniería de la UC, desarrollando nuevas tecnologías para la exploración eficiente de toda la información que se viene en los próximos diez años con la instalación de los nuevos observatorios en nuestro país. BITS

**Rodolfo Angeloni está financiado por el Proyecto Fondecyt #3100029.*

REFERENCIAS

Percival, D. and Andrew T. Walden. (1993) Spectral Analysis for Physical Applications. Cambridge University Press

Bishop, C. (2006) Pattern Recognition and Machine Learning, Springer ISBN 0-387-31073-8.

Bloomfield, P. (1976). Fourier analysis of time series: An introduction. New York: Wiley.

Billères, M., Fontaine, G., Brassard, P., Charpinet, S., Liebert, J., Saffer, R. A., 2000, ApJ, 530, 441.

Cortes, C. and Vapnik ,V. Support vector networks. Machine Learning, 20:273–297, 1995.

Hamilton, J. (1994), Time Series Analysis, Princeton: Princeton Univ. Press, ISBN 0-691-04289-6

Mitchell, T. (1997). Machine Learning, McGraw Hill. ISBN 0-07-042807-7, p.2.

Mills, Terence C. (1990) Time Series Techniques for Economists. Cambridge University Press

N. Roy and A. McCallum. Toward optimal active learning through sam- pling estimation of error reduction. In Proceedings of 18th International Conference on Machine Learning, ICML, pages 441–448, 2001.

N. Cebron and M. Berthold. Active learning for object classification: from exploration to exploitation. Data Mining and Knowledge Discovery, 18(2):283–299, 2008.

Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81-106.

Russell, S.J. and Norvig, P. Artificial intelligence: a modern approach. Prentice Hall series in artificial intelligence. 2010

Y, Baram, R. El-Yaniv, K. Luz. Online Choice of Active Learning Algorithms. JMLR 5:255-291, 2004.

Terrell, D., Mukherjee, J.D., and Wilson, R.E. 1991. "Binary Stars: A Pictorial Atlas", Krieger Publ. Co. (Malabar, Florida).

Wachmann, A.A., Popper, D.M., and Clausen, J.V. 1986, A&A, 162, 62.

Entendiendo la privacidad hoy

Alejandro Hevia
Profesor Asistente, DCC, Universidad de Chile; Ph.D. Computer Science, University of California, San Diego (2006); Ingeniero Civil en Computación, Universidad de Chile (1998). Director del Grupo de Respuesta a Incidentes de Seguridad Computacional, CLCERT.
ahevia@dcc.uchile.cl

En mi época de estudiante de computación, uno de mis profesores tenía colgada en su puerta una caricatura que mostraba dos perros hablando, uno de ellos sentado frente al computador: *"On Internet, nobody knows you are a dog" (o "En Internet, nadie sabe que eres un perro".)* le decía un perro al otro [22]. Para quienes vivieron online esa época, el chiste era claro: en Internet era fácil pretender que se era alguien o algo distinto pues nuestras comunicaciones no tenían mecanismos de autentificación de ningún tipo: tu email o tu página Web era para todos los efectos prácticos indistinguible de la de millones de otras personas.

Hoy en día, la situación no ha cambiado demasiado y todavía no hay autentificación de ningún tipo[1]. Sin embargo, nadie se reiría del chiste anterior. ¿Por qué? El concepto de "quién soy" en Internet ha sido redefinido por el colectivo denominado "Web 2.0", las redes sociales y sistemas de contenido generado por usuarios. "En Internet, tus amigos y seguidores de Facebook, Google y Twitter saben a qué le ladras, cuándo ladras, con quién ladras, qué comes y por dónde paseas. Es irrelevante si eres un perro o no". Probablemente con este título la caricatura ya no sería tan graciosa.

Creemos que el anonimato es prevalente online, cuando de hecho no lo es, o al menos, es muy difícil ser realmente anónimo para un usuario honesto promedio. Es vox populi que en los últimos años, grandes bases de datos con información de cada uno de nosotros han sido amasadas con nuestro conocimiento e incluso con nuestro beneplácito. Esta tendencia de hecho parte fuera de Internet ofreciéndonos beneficios a cambio de un poco de información.

1 A pesar de los deseos de Google, quienes en su red social Google+ inicialmente han restringido el acceso a quienes firmen con "nombres reales" y no seudónimos, algo que ha causado gran controversia y probablemente será cambiado.

Aceptamos la existencia de DICOM como una manera de mejorar el acceso al crédito, o aceptamos revelar nuestra lista de compras a cambio de un beneficio económico (puntos del supermercado). Ya en el mundo online, por un beneficio económico y/o de entretención por ejemplo cedemos gustosamente nuestros datos a cambio de una cuenta gratis en Facebook, donde podemos ver a nuestros amigos y dejarnos ver por ellos. Así, DICOM sabe quiénes somos, a quiénes no les pagamos, y por cuánto. La cadena de supermercados, por otra parte, sabe quiénes somos, qué compramos y cuándo. Facebook, sabe algo más importante: qué tipo de cliente somos, qué queremos y qué quieren nuestros amigos. Y sorprendentemente, toda esta información está en manos de privados. En comparación, el Estado pareciera saber en mucho menos: el Servicio de Impuestos Internos (SII), uno de los mayores recolectores de datos estatales en Chile, sabe esencialmente sólo cuánto ganamos, aunque seguro querría saber aún más. Cierto, los organismos de seguridad estatales legalmente pueden saber mucho más: nos dejamos escanear con rayos X o registrar en los aeropuertos (¡y sorprendentemente aún en las calles!), e incluso nos dejamos escuchar, fotografiar y filmar en la vía pública por el beneficio de la "seguridad" que la policía nos puede proveer en ese ambiente. En suma, aceptamos (incluso aplaudimos) las reglas del juego para obtener los supuestos beneficios. Sin embargo, a muchos todavía les incomoda o incluso se quejan cuando ven reportajes sobre las injusticias que datos incorrectos en DICOM pueden causar, o se quejan de la pérdida de privacidad cuando en las noticias aparecen hackers comprometiendo sitios Web y revelando datos de millones de clientes, o cuando se hacen públicas conversaciones por celular entre congresistas (y más recientemente, entre defendidos y sus abogados), o se publican datos o fotos íntimas de personas sin su consentimiento en portales públicos online. ¿Cuánto realmente queremos tener privacidad? ¿O ya la perdimos? ¿Vale la pena el (aparentemente necesario) costo monetario y social de recuperarla y mantenerla?

¿QUÉ ES PRIVACIDAD?

Entender privacidad es difícil[2]. Según Van der Berg [24], es "probablemente uno de los conceptos más complicados, malentendidos y altamente debatidos en ciencias sociales, en ámbitos legales, filosóficos, y tecnológicos, durante las últimas décadas en el mundo". Esto quizás porque la definición de privacidad no es trivial de entender. Por ello, ruego al lector me permita una sección algo más teórica del tema, con el compromiso de aplicarlo a temas tecnológicos más adelante.

Según Warren y Brandeis [27] privacidad es "the right to be let alone" ("el derecho a ser dejado tranquilo o solo"). Esta definición es notable por dos razones. Primero, por lo adelantada a su tiempo: en ese entonces, a dichos autores les preocupaba la aparición de la fotografía y de las grabaciones como herramientas periodísticas. Y segundo, por tocar un nervio en casi todos los seres humanos: el valor de nuestra individualidad. Sin embargo, es una definición limitada, pues no da demasiadas luces de qué implica, por ejemplo, cuando decidimos dejarla de lado voluntariamente (esto es, cuando a propósito buscamos "dejar de estar solos" y nos contactamos con otros). Una segunda definición la da Burgoon y otros (citado en [17]): "La habilidad para limitar físicamente, vía interacción, sicológicamente e informacionalmente el acceso a mi individualidad o la individualidad de un grupo". Esta definición pone énfasis en las distintas dimensiones de la privacidad: existe en términos físicos (por ejemplo alguien mirando por mi ventana); existe al interactuar con gente (uno la considera al conversar con otra persona); existe en su rol sicológico (expresada en la libertad de tomar decisiones personales, religiosas, de orientación sexual sin presión de otros); y puede dejar de existir cuando la información acerca de una persona es transmitida a otros sin su consentimiento o conocimiento. Una perspectiva distinta la entrega Hildebrant [10] quien la define como "la libertad o carencia de limitaciones irrazonables sobre la construcción de mi propia identidad". Esta definición nos recuerda la íntima relación entre privacidad e identidad. Aquí, una información es privada o sensible si "revela algo respecto a quién soy". Privacidad es algo dinámico y abierto, algo que, tal como nuestra identidad, puede cambiar en el curso de nuestra vida. Finalmente, la definición de Westin [29] se focaliza en un aspecto más preciso, el control del acceso a la información sobre la individualidad de una persona: "La necesidad de individuos, grupos o instituciones de determinar por ellos mismos, cuándo, cómo y hasta qué nivel de información acerca de ellos es comunicada a otros".

Ahora bien, de las decenas de distintas definiciones de privacidad propuestas en las últimas décadas -muchas muy amplias o muy restringidas-, hay una en particular que sobresale en el contexto de tecnologías computacionales: privacidad como "integridad de contexto", propuesta por Nissenbaum [15,16]. Su punto de partida es la existencia de los distintos "mundos" o contextos donde la gente se mueve. *"Al observar la textura de la vida de las personas, vemos que ellos salen y entran, y se mueven dentro de una pluralidad de ámbitos o mundos distintos. Están en sus casas con sus familias, van a trabajar, buscan atención médica, visitan amigos, consultan siquiatras, hablan con abogados, van al banco, asisten a misa, votan, compran, y mucho más. Cada una de dichas esferas, ambientes o contextos involucra, o incluso es definida por, una serie de normas, las cuales gobiernan sus distintos aspectos tales como los roles, esperanzas, acciones y prácticas" [16].* Nissenbaum argumenta que la privacidad de una persona se desprende de su habilidad de compartimentalizar su vida (social), de manera que la información sobre ella, que pudiera ser dañina o vergonzosa cuando es compartida fuera del contexto donde se entregó, se mantenga protegida, circunscrita a las reglas del contexto donde se entregó. Por ejemplo, la gente usualmente no considera privada la información compartida con su doctor en el proceso de obtener atención médica, pero pudiera molestarse si dicha

2 La presentación de las definiciones, historia y varios ejemplos del presente artículo se basan en el excelente reporte titulado "Privacy Enabled Communities" del Primelife Project [18] de la Comunidad Europea.

El problema surge cuando la información es compartida fuera del contexto original, sin preservar las normas (explícitas e implícitas) referentes a cómo se comparte información en ese contexto.

información va a parar a su empleador o a un conocido del colegio de sus hijos. Esta distinción es notable, pues en otras palabras, los datos no son privados o públicos *per se,* sino que lo son sólo respecto a un contexto. No es el contenido de la información, sino el contexto donde fue compartido y, en particular, la audiencia que tiene acceso a dicha información. Esto explica la aparente contradicción vista en personas que comparen información íntima y personal sin mucha preocupación, al tiempo que se sienten profundamente perjudicados cuando ven compartida otra información de ellos aún si esta última no es personal ni sensible. Más aún, entender privacidad bajo la definición de Nissenbaum nos enseña que privacidad es una característica social y no sólo informacional. Compartir información *per se* no es el problema; muchos de nosotros compartimos información con otros todo el tiempo sin estar necesariamente preocupados de las repercusiones de esta conducta en términos de privacidad. Sin embargo, el problema surge cuando la información es compartida fuera del contexto original, sin preservar las normas (explícitas e implícitas) referentes a cómo se comparte información en ese contexto. De hecho, Nissenbaum lo resume en forma excelente: *"Si un ítem de información es considerado apropiado para una situación en particular, típicamente es compartido sin problemas. Más aún, la información puede ser guardada o usada en una situación en particular sin gatillar objeciones de ningún tipo. La gente no objeta el tener que entregarle a los doctores los detalles de su condición médica, o discutir los problemas de sus hijos con los profesores de sus hijos, ni divulgar información financiera a un ejecutivo de cuentas para pedir un préstamo, o compartir con amigos cercanos los detalles de una relación romántica. Para el sinnúmero de transacciones, situaciones y relaciones en las cuales la gente se aboca, hay normas -explicitas e implícitas- que gobiernan cuánta información y de qué tipo es apropiada para ellos. Donde dichas normas son respetadas, diremos que la integridad de contexto se mantiene. Donde no, diremos que la integridad de contexto ha sido violada"* [15].

DE VUELTA A LA PRÁCTICA

Entender privacidad como mantener la integridad de contexto de (por ejemplo) la información asociada a nosotros mismos, puede ayudarnos a entender nuestro comportamiento ante los eventos mencionados al comienzo. Por ejemplo, respecto a los datos en DICOM, la regla implícita (supuesta) es que dichos datos serán fieles representantes de nuestro crédito, que serán preservados (no modificados arbitrariamente), y que serán comunicados apropiadamente (por ejemplo sólo a quienes debieran revisar nuestro crédito y no a otras entidades que aparentemente no debieran necesitarla, como un futuro empleador, o mis amigos o mi doctor). Datos incorrectos en DICOM son una violación a la privacidad en el sentido que vulneran una parte de esta regla, la integridad de los datos en particular. En el caso de las fotos privadas reveladas online, claramente dicha información ha cruzado a un contexto para el cual nunca fueron orientadas (lo cual explica por qué la gente sigue tomándose este tipo de fotos, aún cuando son aconsejados de lo contrario). Dichas fotos tuvieron su contexto, íntimo quizás, en el cual no se consideraban privadas. Y en el caso de los petabytes de información en Facebook, la mayoría de los usuarios implícita o explícitamente creen que ciertas normas se respetarán en el contexto formado por sus amigos. O bien, no entienden cuál es ese contexto. En muchos casos, hay una desconexión con la realidad. Por ejemplo, aunque técnicamente sea posible que un futuro empleador mire las fotos compartidas de una persona buscando trabajo, frecuentemente dicha persona no considera como una violación de su privacidad el compartirlas, pues cree (sin razón) que dichas fotos sólo permanecerán en el contexto (real o imaginado) de sus amigos. Obviamente, esa persona siente su privacidad violada cuando aparecen en el computador del entrevistador o, peor aún, ventilados en una reunión familiar.

PRIVACIDAD VERSUS SEGURIDAD

Lamentablemente, es frecuente ver discusiones donde privacidad es presentada como algo transable por seguridad pública o el bien común. Gobiernos de todo tipo frecuentemente nos quieren convencer que para garantizar la seguridad pública es necesario recolectar datos de los ciudadanos en forma de monitoreos masivos y extensa minería de datos. Se argumenta que los ciudadanos honestos no debieran temer, pues "quien nada hace nada teme".

Éste es un falso dilema (discutido por muchos, entre ellos Solobe [20,21]), pues presupone que privacidad sólo existe para "esconder cosas malas". El argumento es que gente honesta y sana, no oculta nada, sólo aquellos en los límites de la decencia o la legalidad lo hacen. Con ello se justifica plenamente reducir la privacidad de todos para buscar a las "manzanas podridas". Ejemplo de ello es la instalación de cámaras en las calles de las ciudades, para detectar y perseguir delitos, o, por ejemplo, en la instalación de

escáner de cuerpo en los cruces fronterizos del norte para detectar traslado de droga. Sin embargo, pese a que los objetivos puedan ser loables, esta dicotomía es falsa. Primero, si privacidad es integridad de contexto, todos tenemos algo que ocultar (¿por qué no vivimos en casas de vidrio?). No quiero que mi información médica deje la consulta de mi doctor o que el monto de mis ahorros lo sepa mi vecino. Como vimos, qué es privado depende del contexto: hay veces que información muy íntima (la última cirugía plástica de una actriz famosa) no es considerada privada e información muy pública (su número de carnet) sí lo es. La dicotomía entre privacidad e ilegalidad, por un lado, y seguridad y legalidad, por otro, simplemente desconoce la percepción de la privacidad de la gente en el mundo real. Peor aún, tal argumento (privacidad sólo es para quienes quieren ocultar algo malo) no considera aspectos más fundamentales, como la necesidad de su existencia para tener una sociedad realmente democrática y libre. Valorar y respetar la privacidad permite a la gente hablar libremente, sin temor a manifestar ideas contrarias al credo imperante, y finalmente relegarse, si es necesario, a un espacio que pueden llamar privado, donde la intervención (estatal o privada) no es permitida. Por ende, argumentar que la privacidad de los ciudadanos debe ser disminuida en favor de su (supuesta) seguridad pública va en contra de valores intrínsecamente democráticos. En palabras de Kee Hinckley, criticando la falta de seudoanonimato en la red social Google+: *"El foro de discusión pública ya no es la plaza del pueblo, el diario ni la calle. Es aquí, en Internet, y está sucediendo en comunidades como ésta, hosteadas por compañías del sector privado"*.

Por supuesto, el hecho que mis fotos íntimas o el vídeo de la cámara de vigilancia donde aparezco saliendo de una tienda de ropa látex para adultos aparezca publicado en Facebook o Youtube no tiene nada que ver con la validez de mis opiniones políticas. Sin embargo, todos sabemos que tal hecho ciertamente puede provocar un daño serio a mi reputación, indirectamente descalificándome como interlocutor válido. El punto a recordar es que no sólo debemos pedirle a entidades privadas respetar la integridad de contexto (mi privacidad) sino también a entidades públicas. Este tema de la relación entre privacidad y libertades democráticas es largo y probablemente requiera su propio artículo.

PRIVACIDAD Y TECNOLOGÍAS MODERNAS: ¿QUÉ CAMBIÓ?

Las tecnologías de información han posibilitado como nunca la recolección, manipulación, distribución y mantenimiento de información en una escala masiva. Al usar empresas como Facebook, Google o el Servicio de Impuestos Internos como intermediarios en muchas de las acciones de nuestra vida diaria, les hemos permitido recolectar dicha información a escalas sin precedentes. En palabras de Vint Cerf, uno de los creadores de Internet: *"Nunca en la historia de la humanidad hemos tenido acceso a tanta información tan rápido y tan fácilmente"*. Tal información puede agregarse, copiarse, enlazarse, correlacionarse y distribuirse en forma barata y masiva. En comparación, antiguamente, lograr recuperar información confiable – o armar un "dossier"- respecto a una persona era una labor investigativa mayor: el "investigador" debía visitar hospitales, escuelas, oficinas, municipalidades, bancos, casas, iglesias, etc., todos aquellos lugares donde la persona había estudiado, trabajado, interactuado y vivido. En cada lugar, el investigador debía conversar con quien estuviera en control de los datos y convencerlo de lo apropiado de compartirlos, justificando de paso su autoridad para solicitarlos. Luego debía hacer copias manuales (con suerte fotocopias) de los documentos con los datos. Hough [9] (citado en [18]) argumenta que *"por ineficiente que pareciera, el almacenamiento de registros en papel, en ubicaciones diversas, en realidad creaba un colchón de protección, asegurando que los datos no fueran revelados sin un esfuerzo considerable y sólo con una causa justa"*. No sólo servicios "gratuitos" recolectan esta información, también hay casos emblemáticos de empresas que derechamente los venden. Solove en su libro "The digital person" [20] reporta a una compañía llamada "Regulatory DataCorp" (RDC) la cual ha "creado una base de datos masiva para investigar gente que abre nuevas cuentas bancarias" y que en su base de datos la información es recolectada "desde más de veinte mil fuentes distintas en el mundo". Es discutible cuánto de esa información es estrictamente necesaria para

evaluar el riesgo de un potencial cliente y cuánto de ella simplemente es un "dossier" de la persona.

Quizás todo esto fue lo que llevó a Scott McNealy, ex CEO de Sun Microsystems, a argumentar *"you have zero privacy anyway. Get over it" ("tienes cero privacidad de todas maneras. Resígnate.")* ¿Es cierto que hemos perdido toda privacidad? McNealy fue altamente criticado por su postura pues aunque correcta en los hechos, fallaba en la reacción: resignación no es la acción adecuada. Ustedes y yo podemos haber perdido nuestra privacidad online, pero ¿nuestros hijos y nietos deben también perderla?

Paradójicamente, las características que más beneficios han traído a la manera de organizar la información en el mundo son aquellas con el mayor riesgo de perjudicar nuestra privacidad. En el proceso de combinar, mover, copiar y editar información desde lugares dispares es que violaciones de integridad de contexto pueden producirse. De hecho, Nissenbaum [15] distingue dos niveles de problemas: (1) los tipos de violaciones ocurridos en el proceso de mover información de un contexto a otro, y (2) aquellos ocurridos en el proceso de combinar distintos trozos de información. Ejemplos del primer tipo son las clásicas violaciones producidas al copiar información médica o financiera a otros contextos donde su uso no fue contemplado. Situaciones del segundo tipo ocurren, por ejemplo, cuando compañías de seguro solicitan exámenes médicos creando perfiles sobre los clientes con información irrelevante al simple análisis de riesgo, pero que vulneran su privacidad. El problema puede ocurrir también por subestimar las capacidades actuales de "enlazar" datos. Es el caso de Netflix, compañía que en 2007 reveló su base de datos de recomendaciones de películas hechas por más de 500 mil clientes, con la esperanza de obtener mejores sistemas de recomendación. Y aunque dijeron tomar especial cuidado de "anonimizar" los datos (eliminando datos personales y reemplazando nombres con identificadores al azar), Narayanan y Shmatikov, dos investigadores de la University of Texas at Austin [14], mostraron que podían "desanonimizar" a muchos de los clientes simplemente comparando información de rankings y fechas/horas con aquellas disponibles en Internet Movie Database (imdb.com), un sitio de recomendación de películas donde los usuarios sí entregan sus nombres. Éste y otros casos similares (como cuando gente fue individualizada a partir de logs "anonimizados" publicados por AOL en 2006 [19], o experimentos que mostraron que el 80% de los ciudadanos en Estados Unidos pueden ser individualizados a partir de los datos del censo, su código de área, su género y su fecha de nacimiento [23, 8]) mostraron que los mecanismos básicos para anonimizar datos no son suficientes para prevenir violaciones de privacidad.

De hecho, en el contexto de recopilación masiva de datos, hay quienes argumentan la necesidad de recuperar nuestro "derecho a ser olvidado" (ver Solove [20]). Biológicamente, nuestra memoria nos ayuda a recuperarnos y comenzar de nuevo, luego de errores o experiencias traumáticas simplemente dejándonos olvidar dichos eventos. *"Olvidar es la norma, recordar es la excepción"*, pero con herramientas ya disponibles podemos perfectamente terminar en *"olvidar es la excepción y recordar es el default"* [12]. ¿Cuántos de nosotros podríamos vivir teniendo "un registro detallado y públicamente consultable" [21] de nuestras acciones, datos y vida desde nuestra infancia? Tal como argumenta Solove [21], no conviene olvidarnos de las ventajas sociales de poder "comenzar de nuevo" y "partir de cero" al considerar la lista de requisitos de nuestra vida digital.

REDES SOCIALES Y REPUTACIÓN

Un tema aparte lo constituyen los posibles problemas de privacidad (y de reputación) derivados del contenido generado por otros usuarios en redes sociales. La identidad de cada usuario no está sólo definida por información entregada por el mismo usuario (su foto, su nombre, sus gustos) sino por información entregada por otros, que puede perfectamente permanecer inmutable en el tiempo pese a ser incorrecta o, aún siendo incorrecta, perjudicar/humillar públicamente a alguien. Un caso emblemático de esto, es el llamado caso de la *"the dog poop girl" o "niña del excremento de perro"* ocurrido en Corea del Sur en 2005. Allí, una mujer en el metro de Seúl se rehusó a recoger el excremento de su perro, lo cual fue fotografiado por otro pasajero. Rápidamente, la foto empezó a circular ampliamente por las redes sociales en el mundo. Luego, su nombre y detalles personales fueron revelados por otras personas en represalia. Al final, su reputación fue arruinada y abandonó la universidad.

¿Podría haber pasado lo mismo sin el apoyo de la tecnología moderna (redes sociales, Internet)? Los rumores y habladurías han existido desde siempre, pero históricamente su alcance ha sido limitado, típicamente al grupo donde se generan. Es sólo con el advenimiento de las redes sociales de gran escala donde este tipo de casos terminan alcanzando audiencias de millones de personas. Se puede argumentar que el supuesto "anonimato" detrás de un nombre de usuario disminuye la inhibición de su dueño y lo hace menos socialmente conciliador en sus comentarios y críticas, pero estudios muestran que comentarios negativos de este tipo aún surgen cuando el autor realiza sus comentarios en forma pública y completamente identificado[3].

Concluyo esta sección poniendo énfasis en cómo las redes sociales cambian el tipo de problemas de privacidad existentes. Por bastante tiempo, mucho esfuerzo técnico (y legal) fue puesto en desarrollar mecanismos de control de información para limitar la filtración de información almacenada y procesada por parte de empresas y organizaciones. Sin embargo, hoy en día poco de ello es aplicable al escenario social, pues en este contexto, quienes comprometen la privacidad de un usuario no son las empresas ni las organizaciones, sino otros usuarios del sistema. Y aunque mecanismos técnicos que nos permitan "contar un secreto" sin temer a la falta de discreción de nuestro confidente en teoría son posibles [18] (la mayoría derivados de "Zero Knowledge", una técnica criptográfica

3 Quizás porque el medio desconecta al autor de su "víctima", algo mucho menos frecuente en comentarios cara a cara.

bellísima y elegante pero poco práctica), no es claro que sean efectivos: muchos de estos esquemas se basan en evitar "generar evidencia" que soporte la indiscreción (el indiscreto no puede probar su aseveración). Es claramente cuestionable si la falta de evidencia ha disminuido la distribución de información incorrecta (pero "jugosa") alguna vez en Internet.

> Gobiernos de todo tipo frecuentemente nos quieren convencer que para garantizar la seguridad pública es necesario recolectar datos de los ciudadanos en forma de monitoreos masivos y extensa minería de datos.

BIG BROTHER, MANY LITTLE BROTHERS

Hoy dejamos muchos rastros "sin movernos del escritorio". Sitios Web nos monitorean en forma distribuida para darnos el dudoso beneficio de mejores avisos (más focalizados, target advertising) o productos gratis. Pero, ¿es tan así? ¿Somos realmente monitoreados? En palabras de Andrew Lewis: *"Si no estás pagando por algo, entonces no eres el cliente; eres el producto en venta"*. Dado lo extenso de este tema, simplemente le sugiero al lector testearlo por sí solo: para saber quiénes lo siguen diariamente, le recomiendo usar un par de días el plugin "collusion" para Firefox [25]. ¡Los resultados son sorprendentes! El nivel de colaboración entre sitios distintos en Internet sólo con el propósito de seguir usuarios es abismante. Aún con esa evidencia a mano, a mucha gente no le importa. Quizás han comprado la excusa de que tal comportamiento es "el precio de obtener productos gratis". El problema más bien pareciera ser el desconocimiento o la falta de concientización respecto a "las posibilidades técnicas para recolectar, guardar y procesar datos acerca de esta persona" [26]. De hecho, estudios muestran que la mayoría de la gente, aunque expresa preocupación por su privacidad online, lo que entiende por "privacidad" es poco claro. Típicamente se refiere "a temores diversos en la red como por ejemplo encontrarse con virus, troyanos y programas espías, atraer spam o ser atacados por un hacker" (ver Paine [17]). Peor aún, Paine reporta que aún aquellos con mejor entendimiento de las amenazas carecían de las herramientas efectivas para protegerse. De hecho, ¿qué es posible hacer para protegerse hoy? Lamentablemente poco, pues los incentivos económicos están en recolectar información. Por ejemplo, para cada nueva técnica de limitar o borrar las "cookies" usadas para seguirnos, surge una nueva técnica para saltarse tal limitación [11]. Tímidamente, por otra parte, iniciativas legales como "Do-Not Track List" [7,28] o sus mecanismos asociados ("Do-Not-Track headers" [6,13]) han tomado fuerza, pero su efectividad está por verse.

Otros tipos de seguimiento "offline" más clásicos, como por ejemplo con cámaras públicas en las calles, se han masificado en las últimas décadas. Para que decir, el sinnúmero de veces que somos captados digitalmente por turistas o cámaras privadas al movernos por la ciudad. Aquí, en términos de privacidad quizás el tema es algo más claro, pues la pregunta no es simplemente si tal desarrollo es deseable o no, sino si tal desarrollo es conveniente para nuestra sociedad. Por un lado podríamos plantearnos que potencialmente el costo del tracking pudiera ser lo suficientemente bajo como para justificar económica y socialmente los beneficios (búsqueda eficiente en la Web, correo y repositorios de vídeos gratuitos con Google, o disminución del crimen en las calles). Pero, la evidencia muestra lo contrario. Aunque el costo directo es pequeño, el costo indirecto es alto comparado con los beneficios. Por ejemplo, en Inglaterra desde hace unos años se cuestionan si el gasto en cámaras es justificable económicamente en términos de disminución efectiva de delincuencia[4], sobre todo considerando los altos costos de estos sistemas [1]. Por otro lado, ni las políticas públicas ni la ley parecieran hacer mucho para disminuir el problema de los miles de "little brothers" que capturan nuestra imagen (física) sin pedirnos permiso. El tema en particular no es simple. Nos gustaría que limitaran legalmente la captura de imágenes de nosotros en la vía pública, pero tal limitación hoy en día es difícil de asegurar técnicamente. A nadie le gustaría, por ejemplo, ser demandado porque la cámara de seguridad de su casa tomó la foto del vecino al pasar.

Volviendo a las redes sociales, los "little brothers" aquí son conocidos: mis amigos y familiares son posiblemente mi mayor fuente de problemas de privacidad. Hoy en día, son ellos, nuestros amigos, conocidos y familiares quienes revelan la mayor cantidad de información respecto a nosotros. Y aunque algunos mecanismos tecnológicos (criptográficos) permiten limitar quiénes pueden acceder a datos sensibles (como el plugin Scramble! para Facebook [2]), el tema está en su infancia.

PRIVACIDAD Y BIG DATA

Un aspecto interesante en la discusión de privacidad es cómo se inserta en la discusión del "Big Data": esta plétora de "masivas cantidades de información generadas por (y acerca de) gente, cosas y sus interacciones" a las cuales "computines, físicos, economistas, matemáticos, cientistas

4 Y no simplemente de una adaptación de la delincuencia a la ubicación y método de uso de las cámaras.

> Estudios muestran que la mayoría de la gente, aunque expresa preocupación por su privacidad online, lo que entiende por "privacidad" es poco claro. Típicamente se refiere "a temores diversos en la red como por ejemplo encontrarse con virus, troyanos y programas espías, atraer spam o ser atacados por un hacker".

políticos, bioinformáticos, y sociólogos están reclamando desesperadamente acceso" [3]. Y aunque hoy son comunes las discusiones acerca de los pros y contras de usar las grandes bases de datos de Twitter, Google, Facebook, Wikipedia y cualquier otro donde la gente deja rastros, para resolver problemas relevantes a nuestra sociedad (por ejemplo, si "¿la disponibilidad de mejores técnicas de análisis permitirá dar acceso más eficiente a información efectiva a la gente? ¿O será usada para monitorear a manifestantes en las calles?" como señala Boyd [5]) es poca la discusión clara respecto a cómo proceder como sociedad en forma integral en este tema. Según Dana Boyd, "Big Data se ve como pura oportunidad: agencias de marketing como un medio para avisos focalizados más efectivos, agencias aseguradoras como una manera de optimizar sus ofertas, y bancos como una manera de interpretar mejor un mercado complejo". Sin embargo, tal discusión se realiza en un ambiente dinámico, donde "la cantidad de almacenamiento no tiene una cota superior clara y donde las decisiones que se tomen hoy pueden impactar seriamente nuestro futuro" y nuestra privacidad en él [3].

Boyd y Crawford [3] insisten en un punto: hoy debemos preguntarnos qué significa tener acceso a estos datos, quiénes tienen acceso, cómo se establece este acceso y con qué fin. Para ello proponen preguntarse en forma crítica respecto al fenómeno "Big Data", sus supuestos y sus potenciales "biases" o precondicionamientos. Interesantemente, uno de sus cuestionamientos surge del tema privacidad: *"Just because it is accesible doesn't make it ethical" (o "solo porque sea accesible no significa que sea ético (accederlo)")*. Los autores cuestionan la libertad con que investigadores publican análisis de datos que, aunque disponibles online, nunca las personas a quienes se refieren accedieron a tal uso. Como ejemplo mencionan un proyecto de 2006 el cual analizó 1.700 perfiles de Facebook (aparentemente públicos) recolectados desde una "universidad norteamericana del noreste" para hacer un seguimiento de los estudiantes por varios años [30]. El estudio utilizó perfiles "públicamente asequibles" de estudiantes de una universidad (luego identificada como Harvard por terceros). Sin embargo, tales perfiles fueron recolectados por asistentes de investigación (estudiantes) de la misma institución, lo que cuestiona su calidad de perfiles públicos. Además, el proceso de anonimización fue cuestionado y con ello, surgieron quejas por la violación a la privacidad de los participantes del estudio[5]. En particular, Boyd y Crawford critican la falta de cuestionamiento de investigadores respecto a la admisibilidad de usar un conjunto de datos "públicos": *"¿Pueden ser simplemente usados sin pedir permisos? ¿Cuál debiera ser la norma ética que rige tales estudios? Las respuestas no son fáciles pues frecuentemente las violaciones de privacidad no pueden medirse en "daños" específicos al momento de publicarse los datos o incluso dentro de 20 años"* [3]. ¿Debieran los datos de un individuo ser incluidos en un conjunto de datos agregados? Por ejemplo, ¿qué tal si un comentario de un blog de una persona es tomado fuera de contexto y analizado públicamente en un estudio altamente publicitado, sin la persona saberlo? ¿Quién es responsable de que el (o los) dueño(s) de los datos no sean afectados al hacer análisis y publicar el estudio? ¿Qué significa que el dueño de un dato dé su consentimiento, sobre todo si es para cierto contexto? ¿Cambia el contexto dependiendo de los resultados del estudio? (Tal consideración me recuerda una broma televisiva en EE.UU. en la cual se les pedía a hombres jóvenes en la playa, dar un saludo para un vídeo de televisión, a lo cual accedían gustosos, pero luego, cuando se les indicaban que debían mandar saludos al "Gay Channel" muchos de ellos huían). La alternativa de solicitar consentimiento a cada uno de los dadores de los datos utilizados en un estudio/análisis es obviamente impráctica. Sin embargo, no se puede legitimar éticamente su uso simplemente porque los datos son asequibles. *"No porque los datos sean públicamente asequibles significa que fueron pensados para ser consumidos por cualquiera"* [4].

Boyd [5] sugiere a quienes deben analizar datos en el ámbito de "Big Data" los siguientes principios: (1) "Seguridad por 'oscuridad' es una estrategia razonable" (para quienes generan datos), lo que significa que la gente comparte sus datos aún sin mecanismos técnicos de protección efectiva bajo el supuesto implícito que *"nadie grabará públicamente esto y lo ventilara"*. Por lo mismo, Boyd propone respetar tal deseo atendiendo el contexto donde fue hecho. El principio (2) es que *"no todos los datos fueron hechos públicos pensando en que serían publicitados"*, lo cual debiera ser obvio; (3) *"Quienes publican información PII no necesariamente rechazan su privacidad"*, donde PII significa "Publicly Identifiable Information" o Información que identifica públicamente a su donante[6]. El principio (4) *"Agregar y distribuir datos fuera de contexto es una violación de privacidad"* debiera ser obvio para el lector a estas alturas, puesto que se sustenta en la justificación de privacidad como "integridad de contexto". Y finalmente, (5) *"Privacidad no es equivalente a control de*

5 Estos datos inicialmente fueron dados en el contexto que serían asequibles sólo para miembros de la universidad.

6 Segun Boyd, "PII se revela todo el tiempo en redes sociales. Lo que si quieren evitar es 'PEI', 'Personally Embarrasing Information' o Información qué avergüenza personalmente a su donante".

acceso". Nuevamente, este último principio es evidente si diferenciamos la "regla" (quién debe acceder a una información según el contexto) del "mecanismo" (quién puede acceder a la información según el mecanismo técnico empleado, el cual puede fallar o estar mal configurado).

Al lector interesado le recomiendo leer la transcripción de la presentación de Dana Boyd [5].

CONCLUSIÓN

La privacidad, en particular en un mundo digital, es un tema fascinante pues su mera definición no es trivial; sus implicaciones políticas, sociales y culturales pueden ser controversiales, pero su valor es inmenso. Como lo hacía notar en broma un amigo, es uno de los pocos temas donde abogados e ingenieros pueden tener una conversación de genuino interés mutuo. Ello pues la definición de sus límites y consecuencias, y qué mecanismos legales y tecnológicos disponibles pueden ser usados para protegerla no pueden sustraerse de las características de las comunidades mismas donde se intenta preservar. He allí el desafío: nuestra privacidad en el futuro no puede construirse en privado, debemos todos colaborar para lograrla. BITS

REFERENCIAS

[1] BBC report "1,000 cameras 'solve one crime'", Disponible en: http://news.bbc.co.uk/2/hi/uk_news/england/london/8219022.stm, 2009.

[2] F. Beato, M. Kohlweiss, K. Wouters, "Scramble! Your Social Network Data," in Proc. of the International Symposium on Privacy Enhancing Technologies (PETS), 2011.

[3] D. Boyd, K.Crawford, "Six Provocations for Big Data". Disponible en: http://ssrn.com/abstract=1926431, 2011

[4] D. Boyd, A. Marwick, "Social Privacy in Networked Publics: Teens' Attitudes, Practices, and Strategies," paper given at Oxford Internet Institute Decade in Time Conference. Oxford, England. [Citado por Boyd & Crawford, 2011].

[5] D. Boyd. 2010. "Privacy and Publicity in the Context of Big Data" (notas de charla), WWW 2010. Raleigh, North Carolina, abril 2010. Disponible en http://www.danah.org/papers/talks/2010/WWW2010.html

[6] Do Not Track Project, "Do Not Track - Universal Web Tracking Opt Out", Disponible en http://donottrack.us/

[7] Federal Trade Commision, "FTC Testifies on Do Not Track Legislation", Disponible en: http://www.ftc.gov/opa/2010/12/dnttestimony.shtm, 2010.

[8] P. Golle, "Revisiting the Uniqueness of Simple Demographics in the US Population", In proc. of WPES 2006. En http://crypto.stanford.edu/~pgolle/papers/census.pdf

[9] M.G. Hough. "Keeping it to ourselves: Technology, privacy, and the loss of reserve". Technology in Society Vol. 31 (4): 406-413. 2009

[10] M. Hildebrandt. "Technology and the end of law. In Facing the limits of the law", En Claes, W. Devroe and B. Keirsbilck editores. Springer, 2009.

[11] Samy Kamkar, Descripcion del mecanismo "Evercookie", Disponible http://samy.pl/evercookie/

[12] V. Mayer-Schönberger, "Delete: The virtue of forgetting in the digital age.", Princeton University Press, 2009.

[13] Mozilla Do Not Track Project, Disponible en: http://dnt.mozilla.org/

[14] A. Narayanan, V. Shmatikov. "Robust De-anonymization of Large Sparse Datasets" (How to Break Anonymity of the Netflix Prize Dataset). Security &Privacy, Oakland. Disponible en http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf, 2008

[15] H. Nissenbaum, "Protecting Privacy in an Information Age: The Problem of Privacy in Public". Law and Philosophy vol. 17(5-6), pp.559, 596. 1998.

[16] H. Nissenbaum, "Privacy as Contextual Integrity". Washington Law Review. Vol. 79 (119): pp.119-159. 2004.

[17] C.Paine, U-D. Reips, S. Stieger, A.N. Joinson, and T.Buchanan, "Internet users' perceptions of 'privacy concerns' and 'privacy actions'". International Journal of Human-Computer Studies. Vol 65(6), pp. 526-536. 2007.

[18] "D2.4.1 - Final report on mechanisms", Primelife.eu report, Disponible en: http://www.primelife.eu/results/documents/144-241d

[19] SecurityFocus, "AOL search data identified individuals", http://www.securityfocus.com/brief/277, 2006

[20] D.J. Solove, "The digital person: Technology and privacy in the information age". New York. New York University Press.

[21] D.J. Solove, ""I've got nothing to hide" and other misunderstandings of privacy". San Diego Law Review. Disponible en "http://ssrn.com/abstract=998565", vol. 44, pp. 745-772.

[22] Peter Steiner, "On the Internet, nobody knows you're a dog", New Yorker Magazine, publicada 5/Jul/1993, Disponible en http://www.cartoonbank.com/invt/106197, 1993.

[23] L. Sweeney, "Simple Demographics Often Identify People Uniquely". Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh. Disponible desde http://dataprivacylab.org/projects/identifiability/index.html, 2000.

[24] B.van der Berg, R. Leenes (editores), "Privacy Enabled Communities", Reporte de Privacy and Identity Management in Europe for Life, Disponible en: http://www.primelife.eu/images/stories/deliverables/d1.2.1-10.04.23-privacy_enabled_communities-public.pdf, 2011.

[25] Atul Varma, "Collusion Plugin", v.011, Disponible https://secure.toolness.com/xpi/collusion.html

[26] A. Vedder, "Privacy, een conceptuele articulatie", Filosofie & Praktijk, Vol 30(5), pp. 7-19, 2009. Citado en [17]

[27] S. Warren, L. Brandeis, "The Right to Privacy", Harvard Law Review, vol. 4(5), 1890.

[28] Washington Post, "Sen. Rockefeller introduces 'do not track' bill for Internet", Disponible en: http://www.washingtonpost.com/blogs/post-tech/post/sen-rockefeller-introduces-do-not-track-bill-for-internet/2011/05/09/AF0ymjaG_blog.html

[29] A.F. Westin, "Privacy and Freedom", 1st edition, New York, Atheneum, 1967.

[30] M. Zimmer, "On the 'Anonymity' of the Facebook Dataset", Disponible en: http://michaelzimmer.org/2008/09/30/on-the-anonymity-of-the-facebook-dataset/, 2008.

La Web de los Datos

Gentileza: Daniel Hernández.

Claudio Gutiérrez
Profesor Asociado DCC Universidad de Chile. Ph.D. Computer Science, Wesleyan University; Magíster en Lógica Matemática, Pontificia Universidad Católica de Chile; Licenciatura en Matemáticas, Universidad de Chile. Líneas de especialización: Fundamentos de la Computación, Lógica Aplicada a la Computación, Bases de Datos, Semántica de la Web.
cgutierr@dcc.uchile.cl

Daniel Hernández
Estudiante de Magíster en Ciencias de la Computación e Ingeniero Civil en Computación, Universidad de Chile. Entre sus áreas de interés se encuentran la Web, la publicación de datos y el acceso a la información pública.
daniel@degu.cl

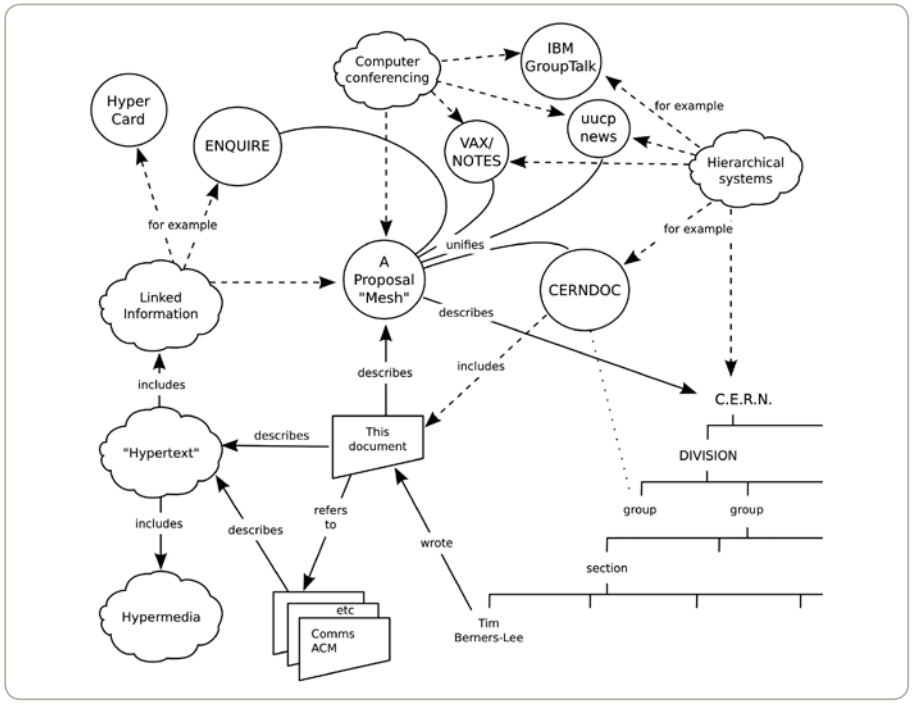
Desde el punto de vista de la información, probablemente la conceptualización más ingenua (pero también más entendible) de la Web sea la de una biblioteca infinita. La idea no es nueva y en 1939 ya Borges la había explicitado en su cuento La Biblioteca Total. *"Todo estaría en sus ciegos volúmenes"*, escribe. De hecho, concebir un espacio universal de información como una generalización de una biblioteca es muy útil. Incluye casi todas las facetas que uno querría que tuviera tal artefacto. Pero tiene un sesgo fundamental: la biblioteca está compuesta de libros, esto es, en términos de la Web, de documentos. Documentos son artefactos producidos por humanos para ser procesados ("consumidos") por humanos. Si uno reemplaza en este modelo el rol que juegan los libros (o documentos) por "datos", lo que se tiene es un modelo abstracto de lo que se llama la "Web de Datos". El cambio parece menor, pero sus consecuencias son impredecibles. Este es el objeto del cual nos ocuparemos en este artículo.

El diluvio de datos. Estudiar la Web de Datos es un tema muy relevante. Permítannos insistir sobre este punto. La gran expectación existente acerca de los inimaginables niveles de producción, disponibilidad y usos de datos (sensores, experimentos, ciencias, estadísticas, redes sociales, etc.) indican que estamos viviendo un cambio fundamental en las prácticas tradicionales de producción, intercambio y procesamiento de la información. La ola de datos fue observada ya hace algunos años por analistas de tecnologías. O'Reilly, en 2005, al tratar la Web 2.0 [1] indicaba que

"los datos son el siguiente Intel". En un nivel más académico, un informe de la comunidad internacional de Bases de Datos [2] advertía que la ubicuidad de "Grandes Datos" iba a remecer las bases de esta disciplina. Szalay and Gray, basados en que la cantidad de datos científicos se duplica cada año, hablaban en 2006 de "un mundo exponencial" [3] y Bell y sus colegas [4] lo llamaron "Diluvio de Datos". Todos se referían al fenómeno del incremento exponencial de volúmenes de datos comparado con el de una década atrás, debido a los avances tecnológicos que permiten capturarlos, transmitirlos y almacenarlos: satélites, telescopios, instrumentos de alto rendimiento, sensores, redes, aceleradores, supercomputadores, etc. Pero el fenómeno no es exclusivo de las áreas científicas. Tendencias similares pueden encontrarse en casi todas las áreas de la actividad humana. Las redes sociales están generando, no sólo grandes volúmenes de datos, sino también redes complejas que piden nuevas técnicas y enfoques para la gestión y procesamiento de datos. Las nuevas tecnologías también han impactado las políticas gubernamentales. Leyes de transparencia e iniciativas de publicación y archivo de datos están imponiendo el mismo tipo de desafíos al sector público [5]. Administrar, curar y archivar datos digitales se ha convertido en una disciplina *per se*. Algunos ya hablan de la "ciencia de los datos" [6]. Este fenómeno está impactando la disciplina de la computación en todas sus dimensiones, desde el nivel de sistemas, arquitecturas, comunicaciones, bases de datos, modelos de programación, ingeniería de software, etc. (ver por ejemplo: [7,8,9]). En todos estos desarrollos, la Web juega un rol central como una plataforma natural donde "viven" y se encuentran estos datos.

En este artículo exponemos sucintamente las iniciativas y tecnologías más relevantes que se han desarrollado para abordar los desafíos del manejo de datos en este nuevo escenario. Presentaremos primero las nociones básicas de la Web. Luego, abordaremos las dos iniciativas más relevantes en estos temas: Linked Data y Open Data. A continuación, presentaremos las técnicas actuales para la publicación y el acceso de datos abiertos. Finalmente, describimos algunas de las herramientas más importantes que se están usando en la Web de Datos.

Figura 1



La primera propuesta de la Web por Tim Berners-Lee. Nótese las ideas subyacentes: datos heterogéneos, usuarios heterogéneos, ausencia de jerarquías, redes, principalmente documentos (tomado de Tim Berners-Lee, Information Management: A Proposal).

BREVE DESARROLLO DE LA WEB

Tim Berners-Lee (TBL en adelante), el creador de la Web, la definió como "un espacio de información compartida a través del cual personas y máquinas se pudieran comunicar" [10]. En otra intervención, insistía que "lo más importante de la Web es que ella es universal" [11]. Veremos que esta universalidad está estrechamente ligada al compartir. No debe ser privativa de una compañía, ni de un gobierno, ni de una organización particular, sino que debe ser compartida por toda la gente alrededor del mundo.

El problema técnico que motivó el primer diseño de la Web, fue desarrollar un espacio para la gente que trabajaba en el CERN, que provenía de diferentes países, con diferentes costumbres, diferentes idiomas; manejando información muy heterogénea, como directorios de direcciones y teléfonos, notas de investigación, informes y mensajes, documentación oficial, etc., y basados en una infraestructura también heterogénea: terminales, servidores, supercomputadores, diversos sistemas operativos, software y formatos de archivos.

Roy Fielding [12], uno de los importantes arquitectos de los protocolos de la Web, resumía estos desafíos así: construir un sistema que debiera proveer una interfaz universalmente consistente a esta información estructurada, disponible en tantas plataformas como sea posible, y desplegada incrementalmente a medida que nueva gente y organizaciones se integren al proyecto.

Administrar, curar y archivar datos digitales se ha convertido en una disciplina *per se*. Algunos ya hablan de la "ciencia de los datos". Este fenómeno está impactando la disciplina de la computación en todas sus dimensiones.

En 2001 TBL [11] recordaba así los desafíos técnicos de tal proyecto:

El concepto de Web integraba diversos y distintos sistemas de información, por medio de un espacio imaginario abstracto en el cual las diferencias entre ellos no existan. La Web tenía que incluir toda la información de cualquier tipo sobre cualquier sistema. La única idea común que amarra todo era la noción de Identificador Universal de Recursos (URI), que identificaba un documento. A partir de allí, una serie de diseños de protocolos (como HTTP) y formatos de documentos (como HTML), que permitían a los computadores intercambiar información, traduciendo sus propios formatos locales en estándares que proveyeran interoperabilidad global.

Resumamos: la arquitectura de la Web se basa en tres pilares

1. **URI** (Universal Resource Identifiers): conjunto de identificadores globales que pueden ser creados y administrados en forma distribuida.
2. **HTTP** (Hyper Text Transfer Protocol): protocolo para intercambiar datos en la Web cuyas funcionalidades básicas son poner datos (put) y obtener datos (get) desde este espacio abstracto.
3. **HTML** (Hyper Text Markup Language): lenguaje para representar información y presentarla (visualmente) a humanos.

De estos tres, los identificadores globales son la base. TBL enfatiza esto diciendo que *"la Web fue diseñada para descansar sobre una especificación: los URI"*. La forma particular que tomaron el protocolo de transferencia (HTTP) y el lenguaje (HTML) fueron soluciones temporales con la tecnología disponible en ese tiempo.

PROTOCOLOS PARA LA WEB

La Web, tal como fue planteada por TBL, podría entenderse como un espacio donde se podría preguntar por URIs y recibir, como respuesta, documentos. Está de algún modo implícito que se espera recibir exactamente aquel documento que es identificado por la URI. No obstante, el protocolo es lo suficientemente abierto para poder implementar otras funcionalidades, por ejemplo, recibir documentos que dependan del usuario. Roy Fielding, de quien hablamos antes, es uno de quienes más ha avanzado en los requerimientos de protocolos Web, es decir, en la definición del comportamiento esperado para interoperar en ella. Por razones de espacio, mencionemos aquí sólo las restricciones que él sugiere en un modelo de arquitecturas que llama REST:

Cliente-servidor. Los clientes deben estar separados de los servidores por una interfaz uniforme. Esto permite modularizar el desarrollo y extensibilidad de las aplicaciones.

Ausencia de estado. Cada pedido de un cliente a un servidor debe contener toda la información necesaria para entender el requerimiento, y no debiera sacar provecho de ningún contexto almacenado en el servidor. El estado de la sesión debiera ser enteramente mantenido en el cliente.

Cacheable. Esto es, que los datos de una respuesta puedan ser implícitamente etiquetados como susceptibles de ser mantenidos o no el en caché. En caso de sí, se da el derecho a reusar esa respuesta para futuros pedidos equivalentes.

Interfaz uniforme. Una funcionalidad central que debiera distinguir la arquitectura de la Web de otras, es el énfasis en interfaces uniformes entre componentes.

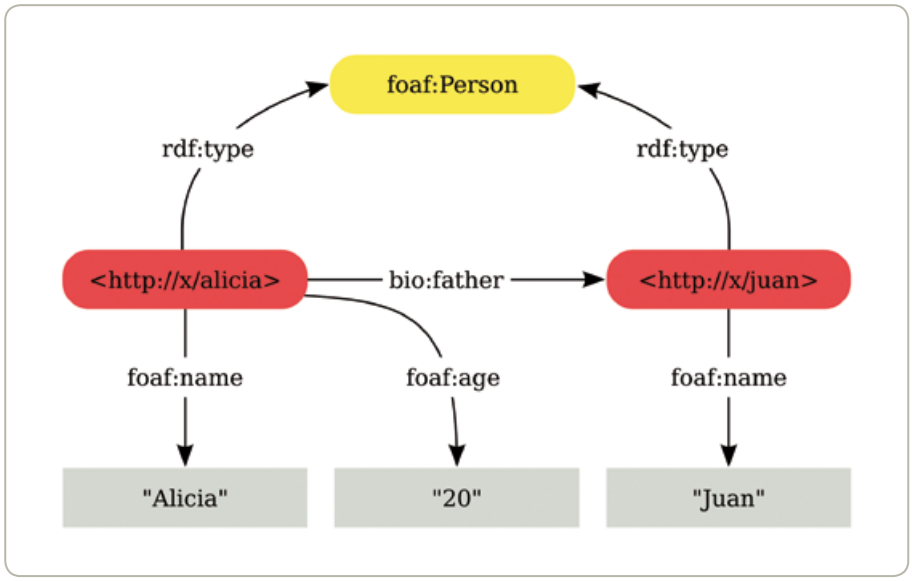
Sistema de niveles. Arquitectura compuesta de niveles jerárquicos, donde cada componente no puede "ver" más allá del nivel inmediato en que está operando.

LENGUAJES PARA LA WEB

Paradójicamente, una de las razones para el éxito de la Web fue la falta de semántica y estructuración del formato de sus documentos, el lenguaje HTML, que surgió más orientado a los elementos visuales que a la codificación de las estructuras de los documentos.

Una segunda generación, XML, permitió definir la estructura de los documentos con mayor precisión, representando los documentos como árboles y agregando reglas que permitieron establecer restricciones en la anidación de los elementos. Las distintas versiones del lenguaje de la Web han ido progresivamente separando la semántica de la presentación, al crear un lenguaje específico para definir la apariencia visual de los elementos, CSS, y retirar atributos que anteriormente permitían definir la apariencia (como @color, @width, etc.). Siguiendo esta tendencia, en las última versión, HTML5, se han incorporado elementos como aside, article, details, menu, nav, header, footer, etc. cuya función es netamente identificar semánticamente la estructura de los documentos.

Figura 2



Ejemplo de grafo RDF. El triple principal en rojo, representa la afirmación "Juan es padre de Alicia". Nótese que además podemos indicar propiedades de cada una de estas personas (rectángulos grises), por ejemplo, nombre y/o edad. Además, podemos incluir información del tipo de objeto de ambos recursos (en este caso son personas, esto es de tipo foaf:Person).

Sin embargo, a pesar del progresivo avance en la separación entre la información y apariencia, para la Web de los Datos esto no fue suficiente, pues el diseño del lenguaje aún tenía en mente el modelo de documento de texto diseñado para ser leído por humanos. Entonces, cabe preguntarse: ¿cuál es el "buen" lenguaje para la representación y el intercambio global de datos? He aquí algunos requerimientos básicos:

1. Que sea suficientemente flexible para describir la mayoría de los tipos de datos (en particular datos, metadatos y conocimiento).
2. Que sea minimalista y eficiente en lo referente a las necesidades de los usuarios y la complejidad de procesamiento.
3. Que pueda escalar en forma distribuida (no centralizada).

La Web Semántica. Hay dos desafíos que motivan una extensión natural de las ideas de la Web a un proyecto que se ha llamado la Web Semántica (en adelante WS): a) si los datos y la información escalan a niveles más allá de la capacidad normal de los humanos (como ocurre hoy día), la única posibilidad de accesarlos, organizarlos y administrarlos es vía automatización. b) El problema del significado de la información: ¿cuál es el significado de cada pieza de información en la Web? Esto tiene que ver fundamentalmente con la semántica y el significado de los conceptos (aún en el mismo lenguaje).

La WS intenta resolver estos problemas basada en la simple idea de organizar la información a nivel planetario. La WS es *"la Web de Datos procesables por máquinas"* escribe TBL. Y esto significa estandarizar significados. Para ello la WS utiliza un modelo de datos que se conoce como Resource Description Framework (RDF) [13] y que está basado en la forma básica de las oraciones, compuestas de sujeto, predicado y objeto. Estas tripletas *(s,p,o)* pueden ser entendidas como fórmulas lógicas binarias del tipo *p(s,o)*.

Un conjunto de tripletas puede ser interpretado como una red semántica, es decir, como un grafo dirigido con nodos y arcos rotulados, donde para cada triple hay un arco rotulado con el predicado y los nodos inicial y final son rotulados con el sujeto y el objeto. Así, la Figura 2 describe dos recursos que representan personas llamadas Alicia y Juan, donde Alicia es hija de Juan.

RDF no sólo describe una estructura de grafos, sino que en él también se definen los conceptos de clase e instancia. En el ejemplo de la Figura 2, los recursos que representan las personas son instancias de la clase foaf:Person.

El otro componente de la WS lo forman la capacidad de establecer reglas que permitan modelar (y validar modelos) y deducir afirmaciones (tripletas) a partir de otras. Para ello se definió The Web Ontology Language (OWL) [14], que es una codificación de la lógica en el lenguaje RDF, diseñado para describir ontologías y asociado a un semántica que define reglas de razonamiento para ellas.

En este punto podríamos detenernos brevemente para señalar una separación entre los caminos del desarrollo de RDF: el de representar datos mediante estructuras de grafos y el de introducir reglas de razonamiento. El primero se enfoca en la idea de una gran base de datos y, como consecuencia natural, requiere de lenguajes de consulta para ella, siendo el más popular SPARQL (una símil de SQL para datos RDF en la Web). El segundo, en cambio, visualiza la información como una base de conocimiento y por ende busca definir reglas para inferir conocimiento a partir de lo ya conocido.

Enmarcados en el compromiso usual entre la expresividad y la complejidad de procesamiento, se han desarrollado varios lenguajes para codificar vocabularios para RDF y por ende las reglas de inferencia que ellos otorgan a los datos expresados. Estos lenguajes pueden ser agrupados, en grueso modo, en tres grupos: a) aquellos con una mínima semántica o sin ella (esencialmente para definir jerarquías de tipos, clases y predicados) [15], b) RDF Schema más algunas extensiones menores y c) OWL, el lenguaje para las ontologías de la Web. No obstante, para enlazar y describir datos, a) pareciera ser suficiente.

LA WEB DE DATOS: LINKED DATA Y OPEN DATA

La Web de Datos es una colección global de datos producidos por la exposición y publicación sistemática y descentralizada de datos (crudos), usando protocolos y lenguajes de la Web.

Sobre la infraestructura de RDF

No es una sorpresa que la noción de la Web de los Datos esté estrechamente relacionada con la WS. Aquí brevemente presentaremos las fortalezas del modelo RDF y los desafíos que se enfrentan para abordar la Web de los Datos.

RDF fue diseñado para facilitar el procesamiento automático de la información en la Web por medio de metadatos. En 1999 la recomendación establecía con claridad: "RDF sirve para situaciones donde la información necesita ser procesada por aplicaciones, en vez de sólo ser desplegada para seres humanos". De este modo, el objetivo principal es la inclusión de información accesible por máquinas en la Web. Pero el diseño de RDF tiene otra consecuencia, su estructura de grafo permite la representación de una amplia gama de datos, abriendo la puerta a la conversión de la Web de los documentos a la Web de los Datos.

El poder de RDF nace de la combinación de dos ideas: a) un modelo flexible para representar tanto datos como sus metadatos de una manera uniforme, en la que ambos compartirían el mismo estatus de objetos de información. b) La estructura de grafos representa naturalmente las interconexiones y relaciones entre los datos. De hecho, esta última característica es la que sustenta el desarrollo de la iniciativa Linked Data.

Linked Data

Entre los proyectos más exitosos que atacan el problema de la ubicuidad de datos en la Web está Linked Data [16,17]. Los autores del proyecto lo definen así [16]:

Linked Data se trata de usar la Web para conectar datos relacionados que no han sido previamente enlazados, o usar la Web para disminuir las barreras para enlazar datos que hoy usan otros métodos. Específicamente, Wikipedia define Linked Data como "una buena práctica recomendada para exponer, compartir, y conectar piezas de datos, información, y conocimiento en la Web Semántica usando URIs y RDF".

La idea es simple: gracias a las tecnologías de la Web, es posible la producción, publicación y consumo de datos (no sólo de documentos) lo que se ha hecho universal. Sacar provecho de esto significa superar uno de los principales problemas hoy: el que estos datos están desconectados unos de otros, impidiendo su aprovechamiento conjunto.

TBL [18] explica como sigue las principales ventajas de Linked Data:

- **Permite conectar diferentes cosas de diferentes fuentes de datos.** El valor agregado de poner datos en la Web estriba en que se los puede consultar en combinación con otros tipos de datos de los cuales uno ni siquiera estaba consciente que existían.
- **Es descentralizado,** permitiendo que cada agencia y persona pueda crear y publicar sus propios datos, sin barreras editoriales, comerciales o administrativas.
- **Uso de estándares abiertos libre de licencias,** significa que nadie, agencias, gobiernos o personas, quedan ligados permanentemente a ningún proveedor.
- **Un círculo virtuoso.** Hay muchas organizaciones y compañías que se motivarán con la presencia de datos para desarrollar sobre ellos diversas aplicaciones y accesos a diferentes grupos de usuarios.

El mismo TBL propuso un test de "cinco estrellas" para la publicación de datos:

1. Ponga su material disponible en la Web (en cualquier formato).
2. Póngalo como datos estructurados (por ejemplo, Excel en vez de imagen escaneada de una tabla).
3. Use formatos no propietarios (por ejemplo, CSV en vez de Excel).
4. Use URLs para identificar cosas, de tal forma que la gente pueda apuntar (y referenciar) a su material.
5. Enlace sus datos con los de otra gente para proveer contexto.

Open Data

Datos abiertos (Open Data) es un movimiento que apunta a facilitar la producción y diseminación de datos e información a escala global. Debido a su relación con los temas que surgen de la discusión de lo "público versus lo privado", el movimiento ha llegado a ser muy influyente en la administración y manejo de la información en gobiernos, bibliotecas y grandes organizaciones.

Podemos definir Open Data de la siguiente manera: "Datos Abiertos es un movimiento cuyos objetivos es desarrollar y difundir estándares abiertos para los datos en la Web".

Por supuesto la gran pregunta es qué significa "datos abiertos". Seguiremos aquí el enfoque metodológico de Jon Hoem en su estudio de comunicación abierta [19], adaptándolo a nuestro ámbito. Hay muchas posibles dimensiones desde donde acercarse a la "apertura" de datos. Tres importantes son: el nivel de contenidos, el nivel lógico y el nivel físico. Para los datos, esto significa informalmente: semántica, tipos de datos y formatos, y hardware.

La gente ligada a datos gubernamentales es quien ha elaborado más a este respecto. Temprano, en 2007, se propusieron ocho principios para datos abiertos [20]. Aunque se refieren a "datos públicos", ellos ofrecen buenos puntos de vista genéricos:

1. **Que sean completos:** todos los datos deben estar disponibles.
2. **Que no estén procesados:** los datos se publican tal como fueron recolectados en la fuente, con el máximo nivel posible de granularidad (sin ser agregados ni modificados).
3. **Que sean actuales:** exponga los datos tan rápido como sea necesario para preservar su valor.

4. **Que sean accesibles:** hacerlos disponibles para el más amplio rango de usuarios y con los más diversos propósitos.
5. **Que sean susceptibles de automatización:** los datos razonablemente estructurados y marcados permiten su procesamiento automático o semiautomático.
6. **Que no haya discriminación:** los datos debes estar disponibles para todos sin necesidad de registrarse.
7. **Que no sean propietarios:** los datos deben estar disponibles en formatos para los cuales ninguna entidad tenga exclusivo control.
8. **Que sean licenciados abiertamente:** los datos no deben estar sujetos a ningún copyright, patente, marca registrada o regulación de secreto de negocio.

Los ocho principios anteriores definen lo que se puede considerar como Datos Abiertos, es decir, no implican que todos los datos deban cumplirlos. Muchas veces pueden existir buenas razones para no hacerlo, como la privacidad y la seguridad.

El impulso dado al desarrollo de modelos de Datos Abiertos ha descubierto varias actividades que eran consideradas como "dadas", o no habían ganado la prominencia que tienen hoy. Particularmente importantes aparecen actividades como preparar datos (para publicación), limpieza de datos, diseño de vocabularios internacionalizados, infraestructura física, disponibilidad de servicios, trazabilidad de origen, y particularmente, temas de licenciamientos y aspectos legales.

PUBLICANDO Y ACCESANDO DATOS ABIERTOS

Ambos proyectos, Linked Data y Open Data, son proyectos relativamente independientes. El primero busca entrelazar información generada y almacenada de manera distribuida y de naturaleza heterogénea con una tecnología apropiada, sitial que de momento es ocupado por RDF. El segundo hace hincapié en que los datos se encuentren disponibles para la ciudadanía, independientemente de la forma en la cual se puedan integrar.

En general las organizaciones se han visto enfrentadas ante la obligación de hacer pública su información. En Estados Unidos esta obligación surgió de una orden emanada desde la Presidencia y en el caso chileno comenzó con la Ley Nº 20.285, sobre el acceso a la información pública. Ante estas ordenanzas, los organismos que las tienen que cumplir se ven enfrentados ante los detalles técnicos y legales, sin poseer marco conceptual que les permita ejecutarlas adecuadamente. La falta de este marco para organizar, preservar y hacer que los datos públicos se mantengan accesibles en el largo plazo ha tenido como consecuencia que mucha información relevante desaparezca o que transcurra tiempo valioso con ella, fuera del alcance de quienes podrían haberla utilizado.

En los tiempos previos al advenimiento de las computadoras plantear metodologías para albergar la información resultaba una tarea sencilla de describir. R. A. Baker resumía las prácticas para mantener ordenados los apuntes de las investigaciones en laboratorios de química hacia 1933 de la siguiente manera [21]:

Dado que la investigación es un esfuerzo organizado para descubrir y una productiva aplicación de hechos, todos los datos obtenidos deben ser adecuadamente ordenados, correlacionados, interpretados y finalmente archivados con el fin de lograr el retorno del esfuerzo invertido. Cada experimento debiera ser titulado claramente y debería limitarse a una materia o variaciones de un sólo factor. El título debería aparecer al inicio de cada página dedicada al experimento. Luego del título inicial debería haber una descripción del problema, seguido del procedimiento, una descripción de los instrumentos, los datos y, finalmente, las conclusiones.

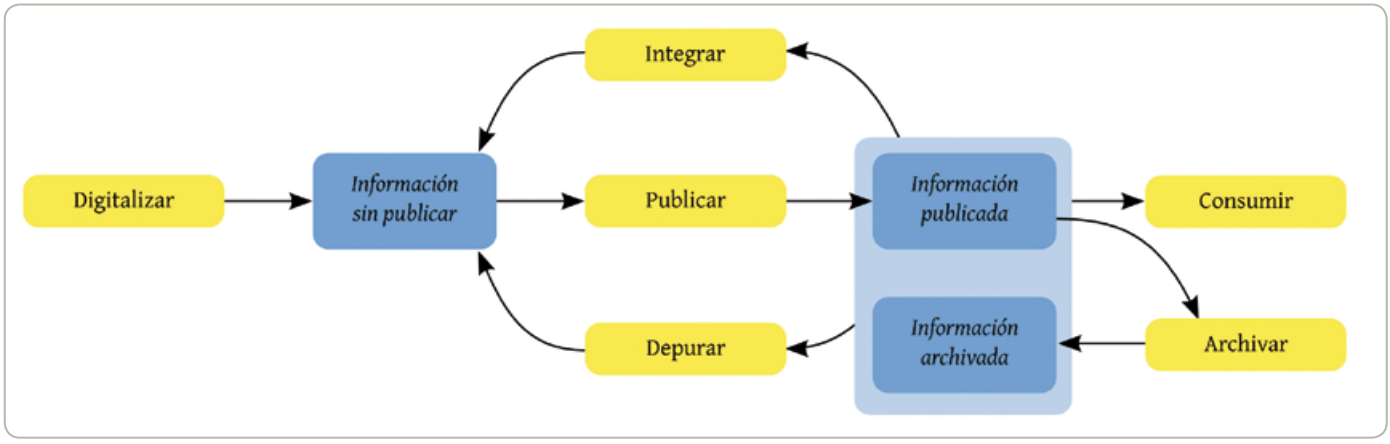
De igual manera las prácticas para mantener los archivos contables de una oficina, los acuerdos, las leyes, los archivos de bienes raíces, la información del registro civil, los registros médicos, etc. se definieron meticulosamente para satisfacer las necesidades de cada área y aprovechar las posibilidades que entregaba la documentación en papel. Sin embargo, con la llegada de las computadoras estas prácticas dejaron de tener vigencia. Las posibilidades de interrelacionar distintas fuentes de datos y de procesar de manera automática los crecientes volúmenes de información impusieron nuevos desafíos a la publicación de datos. En el caso particular de los datos científicos, las buenas prácticas que se ejemplifican en el flujo de datos que resumía R. A. Baker se dejaron de lado. Los medios actuales de comunicación de las investigaciones son los papers, pero estos rara vez van acompañados con referencias a los datos. Lo que es peor, en gran parte de los casos los datos no pueden ser accedidos, pues no siempre son públicos o porque se han desechado.

Sin duda los cambios introducidos por el uso de los computadores requieren hacer un cambio de paradigma en la manera en que tratábamos la información. Ello requiere detenernos a revisar y conceptualizar el proceso de la generación, preservación y uso de los datos en nuestros tiempos digitales.

Datos, datasets, archivos, bases de datos y distribuciones

La primera pregunta que salta frente a nosotros es qué son los datos. En general pareciera haber cierta convención tácita de que los datos deben ser los átomos de la información. Así, pareciera un buen acuerdo concluir que los datos son afirmaciones instanciadas, es decir, expresiones de la forma $a(x,y,...,z)$, donde a es una afirmación sobre los objetos $x,y,...,z$. Esta noción de datos es útil porque podemos llevarla fácilmente a nuestros espacios conocidos de las bases de datos relacionales y al modelo de triples de RDF. En el modelo relacional, cada fila de una tabla t puede ser entendida como una fórmula $t(x,y,...,z)$, donde los parámetros son los valores de las columnas en dicha fila. De manera similar en el modelo RDF, cada triple (s,p,o) puede ser entendido como una fórmula $p(s,o)$.

Figura 3



Acciones y estados de la información.

Con este concepto de datos nos resulta sencillo definir un dataset como un conjunto de datos que puede ser definido por extensión, enumerando todos los datos, o por comprensión, cuando podemos acotarlo de alguna manera aunque luego no podamos enumerar todos los datos. Por ejemplo, el conjunto de datos de todos los nacimientos en Chile durante 2010, es un conjunto que podríamos poner por extensión, mientras que el conjunto de las edades de todos los chilenos no, pues es algo que va cambiando y cualquier enumeración quedará rápidamente obsoleta. En lo siguiente, a los datasets expresados por extensión los llamaremos datos muertos, mientras que a los otros, datos vivos.

La clasificación entre datos vivos y muertos, nos facilita la diferenciación entre bases de datos y archivos. Un archivo es una secuencia de bits que podemos guardar o enviar por la red. En particular un archivo puede codificar un conjunto de datos por extensión, pero no uno por comprensión. Por otro lado, muchos datasets definidos por comprensión corresponden a bases de datos, cuyos contenidos cambian constantemente. Para interrelacionar ambos conceptos, podemos observar que el *dump* de una base de datos es siempre un archivo.

Por último, un concepto introducido por ontologías para catálogos de datos como DCat es el de distribución. Una distribución de datos es un medio por el cual podemos acceder a un dataset. En la Web las distribuciones son identificadas por URIs, que nos permiten descargar el archivo correspondiente a un dataset, cuando éste es expresable por extensión, o acceder a una interfaz que nos permite consultar la base de datos que lo define.

Actores y procesos en la vida de los datos

En una primera aproximación al mundo de los datos podemos suponer dos actores: quienes publican la información y quienes la usarán. No obstante, los roles que encontramos en los participantes son más variados y muchas veces los agentes participan cumpliendo más de un rol. La Figura 3 grafica un modelo algo más detallado de los roles de los actores definidos por sus actos (amarillos) y los estados por los que la información pasa (azules) como resultado de dichos actos.

Actualmente tenemos una pérdida entre la digitalización y el consumo. No todos los datos generados se encuentran disponibles para su consumo. Las preguntas son: ¿dónde se están perdiendo los datos? ¿Por qué y cómo podemos evitar que esto suceda? Refiriéndose específicamente a lo que ocurre con los datos científicos, Michael Witt le llamó a esta pérdida "information bottleneck" [22]. Como se mencionó inicialmente, el aumento en la capacidad de digitalización nos llevó al fenómeno referido como el diluvio de datos. El cuello de botella de información se encuentra entre la información sin publicar y la que está disponible para el consumo, es decir, en el proceso de publicación (ver Figura 3). Este proceso de publicación, también conocido como curación de datos, va desde definir estructuras y modelos apropiados para la información hasta generar identificadores para la información publicada y asegurarse de que ella quede accesible para el consumo. Las tareas de depuración e integración, dibujadas como procesos independientes en la Figura 3, pueden también encontrarse en el proceso de publicación en la medida que se busca agregar valor a los datos a publicar.

Dado el gran volumen de la información disponible para ser publicada, el consumo también presenta un desafío que debe ser facilitado en la publicación de los datos. De este modo, la publicación debe facilitar la automatización de procesos tales como: encontrar fuentes de información, buscar información dentro de ellas, extraer partes, integrar y visualizar.

Integración

Es uno de los mayores desafíos que impone la publicación de datos en la Web. La integración de datos consiste en proveer a los usuarios (o consumidores) una interfaz común para acceder transparentemente a datos dispersos y de naturaleza heterogénea [23]. Por ejemplo, un hipotético servicio que

recoge datos de pronósticos meteorológicos provenientes de la Dirección Meteorológica de Chile (meteochile.cl), los contrasta con un servicio extranjero como The Weather Channel (weather.com) y, además, entrega información sobre la disponibilidad hotelera en las distintas localidades.

Además la integración, es el núcleo de los problemas que se busca resolver con la iniciativa Linked Data. En RDF se proponen las URIs como elemento para identificar recursos y cumplen un rol fundamental en la manera que es posible referirse a recursos comunes desde datasets distintos.

Sin embargo, utilizar URIs no basta, la integración requiere que éstas sean compartidas entre los diferentes datasets. Esto involucra también a aquellas URIs que forman parte de los vocabularios, es decir, aquellas que identifican predicados, clases e instancias de uso común (ejemplo: bio:father, foaf:Person, dbp:Chile).

En vez de definir vocabularios propios, comúnmente se recomienda reutilizar vocabularios existentes con el fin de favorecer la interoperabilidad de la información publicada. No obstante, en algunos casos no resulta posible encontrar vocabularios existentes que se adapten a los datos, ya sea por la inexistencia de vocabularios para describir un área demasiado específica o porque nuestros datos poseen localismos que difieren de los modelos conceptuales que, en su mayoría, son diseñados para culturas que difieren de la nuestra. Aún en estos casos suele ser preferible extender vocabularios existentes a crear vocabularios desde cero.

Las recomendaciones anteriores, se deben en gran medida a que aún no está resuelto el problema de cómo integrar datos expresados con distintos vocabularios. Algunas estrategias para enfrentar este problema son: a) traducir los datos de un vocabulario a otro antes de consultarlos, b) aplicar reglas de deducción al momento de realizar la consulta y c) modificar la consulta de modo que permita trabajar con datos expresados en más de un vocabulario. No obstante, el problema de la integración de datos que usan distintos vocabularios en la Web es un problema abierto.

> Sin duda los cambios introducidos por el uso de los computadores requieren hacer un cambio de paradigma en la manera en que tratábamos la información. Ello requiere detenernos a revisar y conceptualizar el proceso de la generación, preservación y uso de los datos en nuestros tiempos digitales.

Otra barrera a la integración de los datos es la ausencia de un modelo universal de la información. Antero Taivalsaari [24] lo resume brevemente:

Un ejemplo de un concepto que es difícil de definir en términos de propiedades compartidas es "obra de arte". Ya que nadie puede definir límites claros para qué es arte y qué no lo es, no hay ninguna clase general "obra de arte", que comparta propiedades comunes. La definición es subjetiva y depende en gran medida de la situación o del punto de vista.

Algunas personas viviendo cerca del Ecuador no pueden distinguir entre hielo y nieve, mientras los esquimales tienen numerosas palabras para distinguir entre distintos tipos de nieve. Los Dani, de Nueva Guinea, tienen sólo dos términos de colores básicos: mili (oscuro / frío) y mola (luminoso / cálido) que cubre el espectro completo, y tienen gran dificultad para diferenciar entre colores con mayor detalle.

Los lenguajes para definir vocabularios RDF Schema y OWL se fundamentan en la definición de clases y subclases, lo que implica establecer jerarquías entre ellas. Las observaciones de Taivalsaari ponen en duda que tal construcción pueda extenderse a nivel planetario. Un fenómeno similar puede visualizarse en bibliotecología, donde las grandes jerarquías ceden paso a pequeños tesauros funcionales que pueden aplicarse simultáneamente para describir un mismo conjunto de recursos. Siguiendo la estrategia de los pequeños tesauros, Simple Knowledge Organization System (SKOS) es un lenguaje que permite definir esquemas conceptuales para ser aplicados independientemente unos de otros, sin requerir la construcción de una jerarquía única.

HERRAMIENTAS PARA PUBLICAR

Diversas herramientas han surgido a la par con las necesidades identificadas en la práctica de la publicación de datos. La mayoría de los organismos públicos que tomaron el desafío de hacer accesible la información pública a la ciudadanía comenzaron con catálogos de datos, donde los datasets, al igual que los catálogos de documentos, eran tratados como objetos opacos, en los cuales sólo es posible acceder de manera uniforme a ciertos metadatos comunes. Los catálogos cumplen con el objetivo básico de hacer accesibles y referenciables a los datasets, pero aún presentan una deuda: la integración de datos. Es allí donde el modelo RDF entra en juego, proveyendo de herramientas para integrar lógicamente los datos y para consultarlos. Aunque aún quedan temas abiertos, como el balance entre la centralización y la distribución.

Catálogos

La creciente publicación de datos por gobiernos y organismos públicos se ha realizado mayoritariamente en la forma de catálogos de datos. La publicación de catálogos nacionales de datos fue impulsada con el precedente establecido por los Gobiernos de Estados Unidos y Reino Unido, con sus catálogos lanzados en mayo de 2009 y en enero de 2010, respectivamente. En un corto período de dos años ya han surgido numerosos catálogos de gobiernos locales, regionales y nacionales, así como también de organizaciones internacionales como el Banco Mundial y numerosas ONG. Existen varias organizaciones preocupadas de hacer una suerte de metacatálogo, es decir, listar y describir todos los catálogos de datos existentes, entre ellos destacan los de la fundación CTIC, la Open Knowledge Foundation (OKF) y el Resselaer Polytechnic Institute (RPI). En el más reciente recuento, la OKF contabiliza la existencia de 139 catálogos de datos.

En Chile, si bien existen varios organismos públicos que están dejando disponible la información, aún no se ha logrado lanzar un portal que permita un acceso común a todas las fuentes de datos nacionales, por lo que muchas de ellas son desconocidas para la población. Junto con las dificultades de encontrar, la información publicada por la mayoría de los organismos públicos chilenos suele encontrarse en formatos que dificultan su procesamiento automatizado e integración con otras fuentes de datos.

Un catálogo puede entenderse como una colección de entradas describiendo conjuntos de datos, también conocidos como datasets. La descripción de los datasets suele incluir metadatos tales como el nombre, la descripción, las materias tratadas, el origen, la fecha de publicación, las licencias de uso, etc. Entre estos metadatos resultan fundamentales las referencias para poder acceder a los datos. En algunos casos estas referencias son teléfonos o direcciones para consultar por ellos, como en el caso del catálogo de datos geográficos de Chile, mantenido por el Servicio Nacional de Información Territorial (SNIT). No obstante, cuando se habla de catálogos de Datos Abiertos lo esperable es que éstos sean accesibles a través de la Web, ya sea mediante documentos descargables (datos muertos) o servicios que permiten consultar por datos en línea (datos vivos).

La gran aceptación que ha ganado el proyecto Linked Data, ha influido en que algunos catálogos modelen y publiquen la información de los datasets con RDF. Un ejemplo de ello es el catálogo de Australia, donde las páginas del catálogo se encuentran en formato RDFa, una extensión de XHTML que permite marcar datos usando el modelo RDF. Para el catálogo de datos públicos del Gobierno australiano se creó un vocabulario RDF específico para expresar sus metadatos, el AGLS, aunque también existen vocabularios de uso general para catálogos como DCat y VoID. El primero de ellos es aplicable para catálogos donde los datos pueden ser publicados en cualquier medio, mientras que el segundo, es específico para interrelacionar datasets publicados en el modelo RDF, ya sea a través de archivos o servicios de consulta (SPARQL endpoints).

El uso de catálogos para dataset responde principalmente a la necesidad de encontrar fuentes, mencionada al inicio de la sección "Actores y procesos en la vida de los datos" de este artículo, y entregarles identificadores que permitan agregar metadatos a los dataset. Así por ejemplo, el problema de los identificadores de datasets publicados de manera distribuida es resuelto por el proyecto Dataverse, utilizando el Universal Numeric Fingerprint (UNF), un identificador generado aplicando una función sobre el dataset con una muy baja probabilidad de colisionar. No obstante lo anterior, el problema de integrar los datos no es abordado en los catálogos, pues ellos se sitúan en un nivel en el cual los datos son visualizados como objetos oscuros de los que sólo se puede agregar información por medio de metadata.

Como comentamos anteriormente, una de las cualidades que hacen relevante a RDF es su modelo flexible para representar tanto datos como sus metadatos de una manera uniforme, en la que ambos comparten el mismo estatus de objetos de información. Así pues, la siguiente herramienta que describiremos, los SPARQL stores, se enfrentará directamente con el problema de la integración.

SPARQL stores

En general hablamos de RDF stores o triplestores para referirnos a bases de datos orientadas a almacenar y consultar datos en forma de triples RDF. En particular, hablamos de SPARQL stores cuando el lenguaje de consulta es SPARQL.

SPARQL es un lenguaje de consulta basado en patrones, es decir, para obtener un conjunto de recursos que satisfacen ciertas propiedades debemos establecer un patrón por medio del cual estos recursos se encontrarán en el espacio de información sobre el cual queremos buscar. Como el espacio de RDF corresponde a grafos, los patrones serán definidos con grafos. Por ejemplo:

```
SELECT ?a, ?b
FROM <http://x/grafo>
WHERE {
 ?a rdf:type foaf:Person .
 ?b rdf:type foaf:Person .
 ?c rdf:type foaf:Person .
 ?a bio:father ?c ;
 ?c bio:father ?b ;
}
```

Busca a todos los pares de nodos *(?a,?b)* dónde *?b* es el abuelo paterno de *?a*. Los elementos *?a, ?b y ?c* son las variables dentro del patrón que corresponde a lo que se encuentra entre los paréntesis que acompaña al WHERE. Las variables deben ser instanciadas para entregar la respuesta que se pide en el SELECT. Por último, FROM especifica el grafo desde donde deben tomarse los triples que se usarán para hacer calzar el patrón.

La noción de grafo, identificable mediante URIs, permite agregar metadatos a estos grafos, visualizándolos como datasets. Esto es especialmente relevante para manejar la proveniencia (linaje) de los datos, porque en muchos casos éstos podrían provenir de diversas fuentes, con distintas calidades, temática y usos de vocabularios.

A pesar de que los SPARQL endpoints satisfacen la necesidad de consultar e integrar datos, actualmente no eliminan la tensión entre concentrar información localmente para consultarla y publicar distribuidamente. Jeni Tennison explica esta situación [25]:

Lo que no me queda muy claro es cómo esta publicación distribuida de datos puede conciliarse con el uso de SPARQL para consultar. Después de todo, SPARQL no soporta (en la actualidad) la capacidad de realizar búsquedas federadas. De este modo, el uso de SPARQL sobre todos los datos enlazados distribuidos suena como si necesitáramos un triplestore central que contenga todo lo que querríamos consultar.

Publicación de RDF en la Web

Cuando hablamos de publicación de RDF en la Web podemos diferenciar principalmente entre dos estrategias: proveer un servicio de consulta para acceder a los datos directamente, como resulta mediante un RDF store y publicar los datos tal cual, en archivos que sea posible descargar.

La publicación de datos como archivos en la Web tiene dos variantes: a) publicarlos como archivos aparte de los diseñados para la visualización humana y b) usar las mismas páginas Web como soporte para la publicación de datos.

En la primera variante nos encontramos con las diversas serializaciones (sintaxis) de RDF como RDF/XML, N3, Turtle, TriG, TriX, etc. En la segunda variante consideramos lenguajes de marcado como RDFa, eRDF y Microdata, que resultan interesantes, pues permiten publicar datos haciendo pequeñas modificaciones en las plantillas, que generan las visualizaciones de los contenidos.

Por último, los desafíos de almacenar e intercambiar grandes volúmenes de datos llevan a plantear formatos de archivos compactos y que sean capaces de resumir de manera autocontenida lo que contienen, como es el caso de HTC [25]. BITS

REFERENCIAS

[1] T. O'Reilly, What Is Web 2.0, 2005.http://oreilly.com/web2/archive/what-is-web-20.html.

[2] R. Agrawal et al., The Claremont Report on Database Research, 2008. http://db.cs.berkeley.edu/claremont/.

[3] A. Szalay, J. Gray, Science in an Exponential World, Nature, Vol. 440, marzo de 2006, pp. 413–414.

[4] G. Bell, T. Hey, A. Szalay, Beyond the Data Deluge, Science, Vol. 323, marzo de 2009, pp. 1297–1298.

[5] DATA.gov, proyecto de publicación de datos del gobierno de Estados Unidos. http://www.data.gov/

[6] Mike Loukides, What is data science?, 2010. http://radar.oreilly.com/2010/06/what-is-data-science.html

[7] G. Bell, J. Gray, A. Szalay, Petascale Computational Systems: Balanced CyberInfrastructure in a Data-Centric World, Computer, Vol. 39, Issue 1, enero de 2006, pp. 110–112.

[8] M. Stonebraker, S. Madden, D. J. Abadi, S. Harizopoulos, N. Hachem, and P. Helland, The end of an architectural era: (it's time for a complete rewrite), Proc. VLDB '07, 2007. pp. 1150–1160.

[9] No SQL, http://nosql-database.org/

[10] T. Berners-Lee. WWW: Past, present, and future. IEEE Computer, 29(10), octubre de 1996, pp. 69–77.

[11] T. Berners-Lee. Commemorative Lecture The World Wide Web - Past Present and Future. Exploring Universality. Japan Prize Commemorative Lecture, 2002 http://www.w3.org/2002/04/Japan/Lecture.html

[12] R. T. Fielding, Architectural Styles and the Design of Network-based Software Architectures. Doctoral dissertation, University of California, Irvine, 2000. http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

[13] G. Klyne, J. Carroll, Resource Description Framework (RDF) Concepts and Abstract Syntax, W3C Recommendation, 2004. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

[14] D.L. McGuinness, F. van Harmelen, OWL Web Ontology Language Overview, W3C Recommendation, 10 de febrero de 2004, http://www.w3.org/TR/owl-features/

[15] S. Muñoz, J. Pérez, C. Gutiérrez, Simple and Efficient Minimal RDFS. J. Web Sem. 7(3), 2009.

[16] LinkedData Project, http://www.linkeddata.org

[17] Ch. Bizer, T. Heath, T. Berners-Lee, Linked Data - The Story So Far, International Journal on Semantic Web and Information Systems, Vol. 3, 2009, pp. 1-22.

[18] T. Berners-Lee, Linked Open Data. What is the idea?, http://www.thenationaldialogue.org/ideas/linked-open-data

[19] J. Hoem, Openness in Communicaton, First Monday, Volume 11, Number 7, 3 de julio de 2006.http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/1367/1286

[20] Seminar on Open Government Data (Open Government Working Group), 7 y 8 de diciembre de 2007. http://resource.org/8_principles.html

[21] Baker, R. A. In the research laboratory. Journal of Chemical Education, Vol. 10, 1933, pp. 408–411.

[22] M. Witt, Institutional Repositories and Research Data Curation in a Distributed Environment, Library Trends, 57(2), 2009. http://docs.lib.purdue.edu/lib_research/104/

[23] T. Lee, Attribution Principles for Data Integration: Policy Perspectives, febrero de 2002.

[24] A. Taivalsaari, Classes vs. Prototypes - Some Philosophical and Historical Observations, Journal of Object-Oriented Programming, 1996.

[25] Jeni Tennison, Distributed Publication and Querying, blog personal. http://www.jenitennison.com/blog/node/143

[26] J. Fernández, C. Gutiérrez, M. Martínez-Prieto, Compact Representation of Large RDF Data Sets for Publishing and Exchange, ISWC 2010. LNCS 6496, pp. 193–208. Shanghai, China, 7–11 November 2010.

Entrevista
Héctor García Molina

Por Claudio Gutiérrez

Héctor García Molina es uno de los científicos de la computación más importantes. Nacido en México, hoy mexicano-estadounidense, es actualmente profesor de la Universidad de Stanford. Fue director del Departamento de Ciencias de la Computación de la Universidad de Stanford, institución en la que fue profesor de Larry Page y Sergey Brin, creadores y fundadores de Google. Su área de especialización son las Bases de Datos Distribuidas. Y es uno de los autores con mayor impacto en Ciencia de la Computación.

Durante el Workshop sobre Fundamentos de Bases de Datos 2011, en Santiago, que lleva el nombre de Alberto Mendelzon en homenaje a uno de los teóricos más importantes de las Bases de Datos a nivel mundial, y gran amigo de la comunidad Latinoamericana del área, tuvimos oportunidad de conversar con Héctor sobre su trabajo y nuestra disciplina.

I. Los inicios en la disciplina

¿Cómo llegaste a la computación?

Desde que estaba estudiando mi carrera en México me empezaron a interesar las computadoras, aunque estudié ingeniería eléctrica originalmente, ya había computadoras y siempre me fascinaron.

Terminé la carrera de ingeniería eléctrica en México en el Instituto Tecnológico de Monterrey, en la ciudad de Monterrey que ahora tiene diferentes campus en México. Después me fui a estudiar a Stanford en 1974 una maestría, inicialmente en ingeniería eléctrica pero al primer año me di cuenta que lo que me interesaba más era el área de computación.

¿Qué cosas te entusiasmaron para cambiarte al área de computación?

Pues el hecho de que era un poco más matemático, podía chequear las ideas un poco más fácilmente, ver que funcionaban y la verdad es que nunca tuve mucha suerte con los circuitos eléctricos. Construí algún circuito y no funcionaba, pero le movía los alambres y sí funcionaba, y luego tenía más

fuerte con los programas que escribía, que eran más sólidos y funcionaban mejor que lo que hacía en electrónica, así que pensé "mejor me voy por la computación".

¿Cómo elegiste el área particular en que trabajas (Bases de Datos) dentro de la Ciencia de la Computación?

Cómo soy de México originalmente y todavía no había decidido dónde me iba a quedar en el largo plazo, quería estudiar algo que además creía que era un poco más práctico y que podía tener potencial más práctico, y en aquel entonces el área de Bases de Datos me pareció un área que tiene más aplicaciones que algunas de las otras áreas de Computación y entonces pensé que tendría la puerta abierta para regresar a México o para trabajar en compañías, porque todavía no había decidido qué quería hacer. Esa fue una de las razones.

¿Y luego de eso, te entusiasmaste y seguiste un Doctorado?

Sí, después de un año me gustó Stanford. Durante mi carrera no sabía bien de qué se trataba un Doctorado, cuando estudié en el Tecnológico de Monterrey había muy poca gente con Doctorado no se hacía mucha o nada de investigación en aquel entonces. No entendía bien en qué consistía hacer un Doctorado, hasta que llegué a Standford y empecé a entender de qué se trataba la investigación, qué hacía un investigador, qué futuro había allí y eso me empezó a entusiasmar y decidí que me quedaba en el Programa de Doctorado.

¿Con quién hiciste la tesis, quién fue tu advisor y qué temas trabajaste?

Mi asesor fue Gio Wiederhold -quien está jubilado pero todavía está ahí y participa en las juntas de nuestro grupo- y fue en el área de Bases de Datos Distribuidas. En aquel entonces me impresionó mucho uno de los trabajos de Jim Gray sobre control de concurrencia y bloqueo; me pareció interesante y tenía algo práctico pero con buena teoría detrás. Me interesó esa área y ahí trabajé inicialmente.

Terminé mi tesis en 1979. Y aún cuando empecé a trabajar en el área de Integración de Información, pues no existía la Web (fue hasta los noventa cuando empezó), o sea había muchas de las ideas pero no tenían la importancia de la Web.

Cuando terminé el Doctorado me fui a trabajar de profesor en la Universidad de Princeton, donde estuve desde el '79 hasta finales del '91, doce años dando clases. Luego tuve la oportunidad de regresar a Stanford como profesor y empecé en enero de 1992.

II. La investigación

En tu larga experiencia como investigador, eres una de las personas que tiene uno de los altos índices de citación en el área, ¿qué desafíos actuales hay para los investigadores de Ciencia de la Computación? ¿Qué has hecho como investigador para mantenerte en las grandes ligas?

Uno de los factores -no sé si desafíos, retos o factores- importantes es estar bien conectado con los problemas de la actualidad: qué está haciendo la industria, de qué problemas está hablando la gente, porque es una buena fuente de problemas y de ideas, y el trabajo que haces entonces es más relevante, le va a interesar más a la gente si estás trabajando problemas que les interesan a las compañías. Creo que esa es una forma de tener éxito en el campo, estar pendientes, estar yendo a conferencias, de visita a compañías, tener interacciones con otros investigadores para estar al tanto de lo que está pasando.

Trabajaste en Princeton y ahora en Stanford, donde el ambiente intelectual y científico es de primer nivel, ¿cuánto incide eso?

Para mí es sumamente importante estar así, en un ambiente de muchas ideas. Tú sabes que muchos de los proyectos que hemos hecho han empezado ya sea cuando uno de mis alumnos -de los cuales aprendo mucho siempre- o alguna visita viene. Por ejemplo teníamos un programa sobre Data Warehousing (almacenes de datos) hace muchos años y eso empezó cuando Dan Fishman que trabajaba en Hewlet Packard Labs pasó a visitarnos, no me acuerdo en qué contexto fue (quizás en una Conferencia), y nos empezó a conversar sobre su interés en el ramo industrial de Data Warehousing. Entonces con mi colega Jennifer Widom nos sentamos y dijimos "no hay nada ahí, pues debiéramos hacer una copia, pero vamos a pensarlo un poquito más, a lo mejor a hay algo más de lo que está hablando la gente y por eso las compañías dicen que hay problemas, veamos cuáles son los problemas que tienen". Entonces comencé a leer, a estudiar y sí descubrimos que había algunos aspectos interesantes en los cuales podíamos contribuir como académicos. Empezamos el proyecto y fue bien recibido, tuvo alto impacto. Ese es un ejemplo de una idea que empezó a partir del contacto con alguien de una compañía. Un problema real. Pensamos si la gente está empezando a hablar de esto, vamos a ver si podemos hacer algo o no.

¿Qué áreas dentro del mundo de los Sistemas de Información y de las Bases de Datos, piensas que son las de más perspectivas hoy, desde el punto de vista técnico?

Es mucha la cantidad de datos que se está generando. Es un problema muy interesante, hay muchas fuentes de información a través de los sistemas sociales, por ejemplo, instrumentos científicos están generando una infinidad de datos, los cuales no sólo se tienen que almacenar y hacer búsqueda de ellos, sino también explotar, hacer data mining y analizarlos. Hay muchos problemas interesantes para poder canalizar efectivamente todos estos datos, entonces es un área muy interesante. Estamos entendiendo más cómo los humanos -porque estos datos son generados por seres humanos- interactúan, se comunican; estamos aprendiendo mucho

sobre la sicología de los humanos y cómo operamos, cómo evaluamos a otros, cómo se propaga la información a través de las redes, que es un área muy importante y muy interesante.

Otra área que está empezando a interesarnos a nosotros y a otros grupos es sobre el uso de seres humanos en la computación, inspirados por "mechanical turk" y otros, donde uno puede dar trabajos pequeños a gente que por un pago pequeño contesta las preguntas o hace los trabajos. Estamos viendo cómo se organiza un sistema de cómputo o un sistema de bases de datos, por ejemplo, donde parte del trabajo y parte de la información esté en las personas, cómo se puede construir un sistema de ese tipo, híbrido entre computación y humanos, y es ahí donde estamos empezando a investigar.

El tema de este diluvio de datos que hablaba Jim Gray, es un tema que parece que está envolviendo a toda la disciplina de la Ciencia de la Computación…

Y más. Todas las disciplinas en general: medicina, economía… todos están queriendo analizar datos.

De repente ¿no te queda la impresión de que la comunidad de Bases de Datos se ha quedado un poquito atrás en eso?

No, pues es difícil distinguir entre las diferentes comunidades, muy diversas, que hacen Bases de Datos: machine learning, data mining, databases (clásicas), Web, etc. Si vas por ejemplo a una conferencia de KDD (Knowledge and Data Discovery) que es una de las principales en esa área, hay gente de Base de Datos y de Inteligencia Artificial que están trabajando ahí. Entonces no creo que nos estemos quedando atrás.

A veces es difícil identificar cuál es la comunidad de Bases de Datos, porque está desperdigada en diferentes subdisciplinas.

Hace años la gente nos decía *"es que la gente de la comunidad de Bases de Datos se quedó atrás con lo de la Web ¿verdad?"* No nos quedamos atrás, es que hay otras disciplinas que están tomando esos trabajos. Pero un poco para presumir ¿de dónde salió Google? La compañía Google salió de un grupo de Base de Datos, por ejemplo. Así que no nos quedamos tan atrás como comunidad.

III. Políticas científicas

Vamos a hablar de políticas científicas. Me gustaría conocer tu opinión, lo pregunto particularmente sobre la computación chilena, sobre esta relación entre la ciencia y la computación como disciplina nueva, y el respeto -o la falta de respeto- con que otras disciplinas la miran en términos de evaluación, publicaciones científicas y estas cosas. ¿Sientes que la Ciencia de la Computación se ha ganado el respeto dentro de las otras ciencias clásicas?

Pues poco a poco. Pero ese problema no es exclusivamente chileno. Fue un problema en Estados Unidos por muchos años, en universidades batallaba la gente en computación para salir adelante en las promociones, porque ocurría el mismo problema: los que estaban tomando las decisiones, los decanos, los rectores de universidades veían por ejemplo el número de publicaciones en journals o revistas y en computación no existe esa tradición porque estamos en conferencias, en sitios "más informales" según ellos.

¿Cómo ganaron esa pelea en lugares como Stanford, porque esa pelea aquí en Chile todavía la tenemos?

No sé cuál es el secreto, pero en algunas universidades se empezaron a tomar más factores, por ejemplo, la opinión de la gente importante, de los que han tenido más influencia en el campo, más que el número de publicaciones o dónde han publicado. Creo que las universidades donde han surgido mejores departamentos de computación es donde hay administradores que aprecian, entienden el campo y saben evaluar las contribuciones e interpretarlas. Por ejemplo en Stanford para promover a alguien el factor más importante son las cartas de evaluación, se piden como a quince personas, especialistas del campo, que evalúen el candidato y eso cuenta más que el número de publicaciones. Sí vemos el número de publicaciones y en qué sitio son, pero es más importante ese otro tipo de evaluación. Creo que el resto de las universidades está cambiando, tendiendo a evaluar la computación en forma diferente.

Hay otro tema que empieza a aparecer en la academia y es esta relación en que el científico más independiente, más puro, se está mezclando mucho con la aplicación y de alguna manera hay un tema económico que tiene que ver con la innovación, con la ligazón con la empresa; aquí alguna gente reclama que esto le hace perder el rol al académico clásico, ¿qué piensas de ese fenómeno?

Me parece que es bueno, como dije inicialmente, tener nexos con el mundo real porque los problemas que generan son interesantes y es la clave para tener impacto, que es una palabra que se usa mucho en las evaluaciones en Stanford y otras universidades. Entonces es más fácil tener impacto si uno está trabajando en problemas reales, donde hay gente esperando la solución. Ahora no es bueno irse al extremo y estar trabajando demasiado solamente en problemas que van a servirle a la industria, porque ésta muchas veces quiere soluciones inmediatas, problemas a corto plazo. Lo que tienen que hacer los académicos es tener los problemas, escoger cuáles son los a más largo plazo en qué pueden contribuir los académicos y no preocuparse tanto por los a corto plazo, porque muchas veces las compañías quieren software, quieren algo inmediato

y muchas veces hay que decirles que no, esa es la clave para poder identificar los problemas que vale la pena atacar.

Y por otro lado, ¿qué opinas de esta competencia casi demencial por el paper que existe hoy? ¿A lo largo de tu carrera notas que ha cambiado, que hay más énfasis en la publicación?

Sí, ha cambiado bastante en los ramos de computación la presión por publicar más y más artículos. Lo veo, por ejemplo, en el número de publicaciones de los que están solicitando empleo. Hace años los currículums llegaban con dos o tres publicaciones y era alguien muy bueno, ahora si no tiene diez o veinte publicaciones alguien que se está recibiendo entonces no es tan bueno.

¿Es parte del fenómeno que estamos viviendo o ves alguna salida a esto?

La verdad no sé cuál es la solución, porque sí hay demasiada presión, demasiadas publicaciones. A mis alumnos trato de decirles que lo importante no es el número de publicaciones que deben tener si no que la gente lea, que tengan impacto, no nada más tratar de sacar un montón de artículos, mejor ir más lento pero tratar de sacar algo que valga la pena. Pero es difícil para ellos aceptar mis sugerencias porque ven que todo el mundo está publicando muchas cosas, entonces a veces les tengo que decir que no manden ese artículo, que no está listo, va a ser contraproducente, que si se lo publican es peor porque no va a ser bueno y a lo mejor van a agarrar mala fama.

¿Cómo ves desde Estados Unidos cómo ha evolucionado la computación en Latinoamérica?

Mis comentarios son más sobre México porque conozco más, no es tanto de Chile en particular, y creo que se aplica a Latinoamérica, pero no estoy seguro.

Parte del problema creo que es que en México, o en parte de Latinoamérica, no hay una cultura de apreciar los estudios de graduados y las ciencias. En México hay mucho más énfasis en tener un trabajo y ganar dinero, más que en tener una trayectoria de investigador, de científico. Por ejemplo, veo en Stanford que nos llegan solicitudes para el Programa de Doctorado de todas partes del mundo y son los mejores estudiantes del mundo que están solicitando, pero hay patrones muy marcados: de ciertos países nos llegan muchas solicitudes y de otros muy pocas, en general de Latinoamérica llegan muy pocas y he chequeado con mi colegas en otras universidades similares y también hay muy pocos solicitantes ¿y por qué?, cada vez que voy a México o a otro país de Latinoamérica trato de hablar con los estudiantes de por qué no quieren estudiar un Doctorado y me preguntan "para qué, para qué voy a estudiar un Doctorado si no me voy a hacer rico, si luego regreso a mí país y no hay puestos de trabajo en el área", así es que no tienen interés. Por otro lado mis colegas en la Escuela de Negocios de Stanford tienen muchas solicitudes de Latinoamericanos. Todos los latinoamericanos aparentemente quieren sacar un máster en administración de negocios, no quieren estudiar computación o ciencias. Estoy exagerando, pero sí es una tendencia.

Tengo la impresión de que aquí en Chile hay un poco más de aprecio a la academia que en otras partes de Latinoamérica, así es que están mejor que en otras partes. Es lo que he visto.

¿Qué mensaje le darías a los profesores y a los investigadores en Ciencia de la Computación en Chile, en dos temas: en el desarrollo de área y en cómo contribuir a la computación desde Latinoamérica?

Hoy en día hay muchas oportunidades de contribuir fuera de los centros principales, porque a través de las redes y de la Web, la información está accesible, los artículos que ustedes pueden buscar en Chile son los mismos que puedo buscar yo en la Web, así que se ha emparejado mucho más el campo, es igual de difícil o fácil trabajar aquí que en otras partes. Ese es un punto importante.

Como sugerencia, tratar de buscar problemas interesantes; ir a conferencias en otras partes del mundo; ir de visita a otras universidades; si tienen oportunidad pasar tiempo en Europa o en Estados Unidos; no preocuparse nada más por publicar muchos artículos sino tratar de tomarlo un poquito más lento y desarrollar las ideas que van a tener gran impacto, porque antes de lanzarse a escribir un artículo les digo a los alumnos que vamos desarrollar tres o cuatro ideas de posibles artículos, explorarlas y de esas cuatro escoger cuál suena más interesante, pero no irse nada más con la primera que se les ocurra.

En cuanto a trabajar solo o no, depende mucho del área y del estilo de cada quien, pero aún si uno trabaja solo es bueno tener colegas con los que se pueda conversar, aunque estén trabajando cosas diferentes, pero con mis alumnos aprendemos mucho cuando tratamos de explicar una idea, el sólo hecho de explicarla y tratar de contestar las preguntas muchas veces le hace ver a uno que a lo mejor esto ya no es tan bueno o hay un problema que aún no se había visto. Creo que es muy importante estar constantemente intercambiando ideas con otros aunque el trabajo sea de uno y los otros estén nada más dando opiniones. O aunque no hagan nada los otros son como el siquiatra: se sienta ahí y deja al paciente hablar. La ventaja es que uno al hablar y explicar entiende mejor las cosas, por eso es muy importante tener colegas que lo escuchen.

¿Qué consejo darías a los estudiantes que están iniciándose en la computación?

Que busquen problemas que los apasionen. Trabajar en un problema que los apasione, que les interese, es lo más importante. BITS

Latin American Theoretical INformatics (LATIN 2012)

Abril 16-20, 2012, Arequipa, Perú.
http://latin2012.cs.iastate.edu

Esta conferencia internacional, usualmente de cinco días, ocurre cada dos años, en marzo o abril. En la actualidad se realiza cada dos años. Sus versiones previas han tenido lugar en Sao Paulo, Brasil (1992), Valparaíso, Chile (1995), Campinas, Brasil (1998), Punta del Este, Uruguay (2000), Cancún, México (2002), Buenos Aires, Argentina (2004), Valdivia, Chile (2006), Buzios, Brasil (2008), y Oaxaca, México (2010). Su temática principal es la Teoría de la Computación ("Theoretical Computer Science"). La conferencia se ha constituido en el evento científico Latinoamericano más importante en la referida temática y ha alcanzado un meritorio reconocimiento internacional.

La próxima versión de LATIN tendrá lugar en la Universidad Católica San Pablo, en Arequipa, Perú. Habrá una sesión especial dedicada a la celebración de los cien años del natalicio de Alan Turing (reconocido como uno de los padres de la computación) y otra sesión en honor al recientemente fallecido Philippe Flajolet (uno de los fundadores del área de análisis de algoritmos). Junto con la conferencia se realizará la 1st Latin American Theoretical Informatics School, dirigida a estudiantes de posgrado y alumnos avanzados de pregrado. Además, durante el próximo LATIN, se entregará por primera vez el premio "Imre Simon Test-of-Time award" que reconoce las publicaciones más influyentes aparecidas hace al menos diez años en LATIN.

La parte principal de la conferencia consta de las presentaciones de las publicaciones aceptadas, complementadas por un grupo de charlista invitados, que en la próxima versión de LATIN se compone de los profesores Scott Aaronson (Massachusetts Institute of Technology), Martin Davis (New York University), Luc Devroye (McGill University), Marcos Kiwi (Universidad de Chile), Kirk Pruhs (University of Pittsburgh) y Dana Randall (Georgia Institute of Technology). En particular, Scott Aaronson y Martin Davis celebrarán con sus charlas el "Alan Turing year" y Luc Devroye las contribuciones científicas de Philipe Flajolet.

La selección de los trabajos a ser presentados en LATIN'2012 estará a cargo del siguiente comité de programa, que reúne a destacados expertos del área:

- R. Baeza Yates, Yahoo!
- N. Bansal, IBM
- J. Barbay, U. de Chile
- M. Bender, Stony Brook U.
- J. R. Correa, U. de Chile
- P. Crescenzi, U. Firenze
- M. Farach-Colton, Rutgers U.
- C. G. Fernandes, U. Sao Paulo
- D. Fernandez-Baca (Chair), Iowa State U.
- G. Fonseca, Unirio
- J. von zur Gathen, U. Bonn
- J. Koebler, Humboldt U.
- Y. Kohayakawa, U. Sao Paulo
- S. R. Kosaraju, Johns Hopkins U.
- R. Kumar, Yahoo!
- G. Manzini, U. Piemonte Orientale
- A. Marchetti-Spaccamela, U. Roma
- C. Martínez, UPC Barcelona
- E. Mayordomo, U. Zaragoza
- L. Moura, U. Ottawa
- J. I. Munro, U. Waterloo
- A. Oliveira, U. Tecnica Lisboa
- L. Rademacher, Ohio State U.
- I. Rapaport, U. de Chile
- A. Richa, Arizona State U.
- J. Sakarovitch, CNRS/ENST
- G. Salazar, U. San Luis Potosí
- N. Schabanel, LIAFA U. París
- R. I. Silveira, UPC Barcelona
- M. Singh, Princeton U.
- M. Strauss, U. Michigan
- W. Szpankowski, Purdue U.
- J. Urrutia, UNAM
- E. Vigoda, Georgia Tech
- A. Viola, U. de la República

Las fechas importantes para el envío de trabajos son:

Fecha límite de envío:
23 de septiembre de 2011

Notificación de aceptación: 25 de noviembre de 2011

Versión final:
16 de diciembre de 2011

Los trabajos deben estar escritos en LaTeX, en estilo LNCS y en inglés. Deben tener una extensión máxima de doce páginas y ser sometidos a través de EasyChair (http://www.easychair.org/conferences/?conf=latin2012).

En particular, se buscan artículos originales en teoría de la computación, incluyendo las siguientes áreas (entre otras): algoritmos (de aproximación, en línea, aleatorizados, teoría de juegos algorítmica), teoría de autómatas, teoría de códigos y compresión de datos, combinatoria y teoría de grafos, complejidad computacional, álgebra computacional, biología computacional, geometría computacional, teoría de números computacional, bases de datos, recuperación de la información, estructuras de datos, Internet y la Web, lógica en Ciencias de la Computación, programación matemática, teoría del aprendizaje computacional, reconocimiento de patrones, computación cuántica, estructuras aleatorias, computación científica.

Los artículos serán publicados en formato electrónico en la serie Lecture Notes in Computer Science (LNCS) de Springer-Verlag. Habrá un volumen especial en alguna revista de corriente principal donde aparecerán artículos seleccionados.

Desde 1975 desarrollando la Ciencia de la Computación en Chile

DOCENCIA | INVESTIGACIÓN | INNOVACIÓN

fcfm

WWW.DCC.UCHILE.CL

dcc
Departamento de Ciencias de la Computación
UNIVERSIDAD DE CHILE

REVISTA
BITS de Ciencia
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

UNIVERSIDAD DE CHILE

Ciencias de la Computación
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

www.dcc.uchile.cl/revista

revista@dcc.uchile.cl